4th Annual Sequencing, Finishing, Analysis in the Future Meeting

Santa Fe, New Mexico

May 27th - 29th, 2009

Los Alamos NATIONAL LABORATORY EST. 1943

JGI

Office of Science U.S. DEPARTMENT OF ENERGY

Contents

The 2009 "Sequencing, Finishing, Analysis in the Future" Organizing Committee:

** Chris Detter, Ph.D., JGI-LANL Center Director, LANL*
** Johar Ali, Ph.D., Technology Development Team Leader, OICR*
** Ruby Archuleta, Genome Administrative Assistant, LANL*
** Patrick Chain, Finishing Coordinator / Group Leader, LLNL*
** Michael Fitzgerald, Finishing Manager, Broad Institute*
** Bob Fulton, M.S., Sequence Improvement Group Leader, WashU*
** Darren Grafham, Finishing Coordinator, Sanger Institute*
** Hoda Khouri, M.S., Staff Scientist, NCBI*
** Alla Lapidus, Ph.D., Microbial Genomics Group Leader, LBNL-JGI*
** Donna Muzny, M.S., Director of Operations, BCM*

***Major Sponsors:* Roche Diagnostics, Covaris, OpGen, illumina, and Integrated DNA Technologies….Thank You!!!**

SameDay® Oligos

Order online by 2:00 P.M. ET and receive your SameDay® Oligos the next business morning via priority shipping.

- 2-OD guarantee (sufficient for > 250 PCR reactions)
- Deprotected & desalted
- Lyophilized
- Available within the U.S. and Canada
- 15-45 bases

Order today at www.idtdna.com

IDT® INTEGRATED DNA TECHNOLOGIES

INNOVATION AND PRECISION IN NUCLEIC ACID SYNTHESIS

www.idtdna.com

ISO 9001:2000 FM88954 ISO 13485:2003 FM 513219

05/27/2009 - Wednesday				
Time	**Type**	**Abstract #**	**Title**	**Speaker**
730 - 830am	Breakfast	x	**La Fonda Breakfast Buffet** (French "Texas Toast", Fluffy scramble eggs, Grilled breakfast potatoes, Applewood smoked bacon, breads, and fruit, etc.)	x
830 - 845	Intro	x	Welcome Intro	**Mary Neu**
x	Session Chair	x	Session Chair	Chair – Donna Muzny
845 - 930	**Keynote**	**FF0086**	**Fulfilling the Promise of a Sequenced Human Genome**	**Eric Green**
930 – 955	Speaker 1	FF0124	Defining Genome Project Standards in a New Era of Sequencing	**Patrick Chain**
955 – 1020	Speaker 2	FF0026	Human Microbiome Project Finishing	**Jessica Hostetler**
1020 – 1050	Break	x	Beverages provided	x
1050 – 1120	Speaker 3	FF0025	Applying Single Molecule Real Time DNA Sequencing	**Steve Turner**
1120 – 1150	Speaker 4	FF0146	Enabling True Biology with Helicos Single Molecule Sequencing	**Patrice Milos**
1150 – 120pm	Lunch	x	**Coronado Lunch Buffet** (Char-grilled chicken breast with barbecue-chipotle vinaigrette, Pan-seared rainbow trout fillet served w/ smoked yellow pepper coulis, Roasted garlic mashed potatoes & seasonal vegetables, etc.)	x
x	Session Chair	x	Session Chair	Chair – Johar Ali
120 – 140	Speaker 5	FF0038	Sequencing Complex Genomes Using Next Gen Platforms	**Tim Harkins**
140 – 200	Speaker 6	FF0093	SOLiD™ Finishing - High Throughput Sequencing, Assembly and Analysis	**Michael Rhodes**
200 – 220	Speaker 7	FF0132	Genomics and Expression	**Haley Fiske**
220 – 345	**Panel Discussion**	x	**Next Generation Sequencing Technology Panel Discussion**	Chair - Bob Fulton
345 – 400	Break	x	Beverages and snacks provided	x
400 - 500	**Tech Time Talks**	FF0013 FF0027 FF0033 FF0016	- A Software System for Validating and Orienting Sequence Contigs Using Optical Maps, - Sequencing & Finishing Pooled BAC Clones at WTSI with New Sequencing Technologies, - Whole-Genome Resequencing by Hybridization of Hemorrhagic Fever Viruses, - Relative Positioning of Scaffolds : a Challenge With New Sequencing Technologies	**Adam Briska,** **Darren Grafham,** **Sofi Ibrahim,** **Valerie Barbe**
500 – 700	**Posters - even #s**	x	**Poster Session**	x
600 - 800pm	**Meet & Greet Party**	x	**Meet & Greet Party - sponsored by Roche --- Food & Drinks**	**Sponsored by Roche**

05/28/2009 - Thursday				
Time	**Type**	**Abstract #**	**Title**	**Speaker**
730 - 830am	Breakfast	x	**Santa Fe Breakfast Buffet** (Scrambled eggs with a choice of three accompaniments on the side -chilaquiles with green chile and cheese, chorizo sausage and roasted green chile, Grilled breakfast potatoes, applewood-smoked bacon and warm flour tortillas, assorted breads and fruits, etc.)	x
830 - 845	Intro	X / FF0156	Welcome Back + **X prize announcement**	**Chris Detter / Larry Kedes**
x	Session Chair	x	Session Chair	Chair – Mike Fitzgerald
845 - 930	**Keynote**	**FF0087**	**Sequencing Complex Regions of the Genome: Disease & Evolutionary Impact**	**Evan Eichler**
930 - 955	Speaker 1	FF0125	Tools for Managing and Comparing Assemblies	**Deanna Church**
955 - 1020	Break	x	**Beverages provided**	x
1020 -1045	Speaker 2	FF0140	Assembly and Finishing of Small and Large Genomes	**Jim Knight**
1045 -1110	Speaker 3	FF0047	ALLPATHS: Assembling Large Genomes with Short illumina Reads	**Sante Gnerre**
1110 - 1135	Speaker 4	FF0130	Adapting Celera Assembler to 454 Platforms and Automated Finishing	**Sergey Koren**
1135 - 1200	Speaker 5	FF0090	AutoSeq: Auto-Assembly Pipeline in a Small DNA Sequencing Core Facility	**Jan Kieleczawa**
1200 - 130pm	Lunch	x	**New Mexican Lunch Buffet** (Pork tenderloin achiote-rubbed and char-grilled with tomatillo-chipotle sauce, your choice of either Chicken or Cheese enchiladas with red or green chile, etc.)	x
x	Session Chair	x	Session Chair	Chair – Patrick Chain
130 -150	Speaker 6	FF0024	Towards the $1000 Genome - the Impact of New DNA Technologies on Finishing	**Niranjan Nagarajan**
150 -210	Speaker 7	FF0035	To 'Finish' or Not to 'Finish'-the $64K Question in Genomics - Data Mining Using 454 Draft Sequences	**Shanmuga Sozhamannan**
210 – 230	Speaker 8	FF0066	New Technology Drafts: Production and Improvements	**Alla Lapidus**
230 – 250	Speaker 9	FF0097	Microbial Finishing at BCM-HGSC	**Shannon Dugan**
250 – 310	Speaker 10	FF0081	Prescreening of Data from GAii Sequencer Resulting in High-Quality Results in Genome Finishing	**Cliff Han**
310 – 330	Break	x	**Beverages and snacks provided**	x
330 – 430	**Tech Time Talks**	FF0110 FF0104 FF0096 FF0071	- Throughput to Insight... and Tackling the Messy Bits Between, - Sequence Enrichment Applying RainDance Technology, - Next Generation DNA Polymerases and Reverse Transcriptases, - Optimizing DNA Shearing for Next-generation Sequencing	**Jarret Glasscock,** **Vincent Magrini,** **David Mead,** **James Laugharn**
430 - 630	**Wine & Cheese**	x	**Beverages, Wine & Cheese provided - sponsored by OpGen**	x
430 - 630	**Posters - odd #s**	x	**Poster Session with Wine & Cheese - sponsored by OpGen**	x
630 - bedtime	on your own	x	Dinner and night on your own - enjoy	x

05/29/2009 - Friday				
Time	**Type**	**Abstract #**	**Title**	**Speaker**
730 - 830am	Breakfast	x	**Healthy Start Breakfast Buffet** (Scrambled Eggs on side tomatoes, scallions and spinach, Turkey sausage links, Assorted chilled fruit juices, Platter of freshly sliced seasonal fruit, Assorted and bran muffins with butter, Granola and oatmeal served with low-fat milk, Individual assorted fruit yogurts, etc.)	x
830 - 845	Intro	x	Welcome Back	**Chris Detter**
x	Session Chair	x	Session Chair	Chair – Alla Lapidus
845 - 930	**Keynote**	**FF0088**	**A Common Framework for Multiple Sources of Bacterial Annotation**	**Owen White**
930 – 950	Speaker 1	FF0005	BioHDF: Toward Scalable Bioinformatics Infrastructures	**Todd Smith**
950 – 1010	Speaker 2	FF0008	Using Consed and Cross_match in Resequencing Projects	**David Gordon**
1010 – 1030	Break	x	**Beverages and snacks provided**	x
1030 – 1050	Speaker 3	FF0126	Medicago truncatula Resequencing of 384 Lines	**Joann Mudge**
1050 – 1110	Speaker 4	FF0084	Performance Comparison of Multiple Genome Partitioning Technologies	**Jon Armstrong**
1110 – 1130	Speaker 5	FF0141	The IMG Systems for Comparative Analysis and Annotation of Microbial Genomes and Metagenomes	**Victor Markowitz**
1130 – 1150	Speaker 6	FF0022	GenePRIMP: Improving Microbial Gene Prediction Quality	**Amrita Pati**
1150 – 1210	Speaker 7	FF0137	Automated Microbial Genome Annotation: the Current State and Future Challenges	**Miriam Land**
1210 - 1230	Closing Discussions	x	**Closing Discussions -** discuss next year's plans	Chair - Chris Detter
1230 - 200pm	Lunch & Close of meeting	x	**La Fiesta Plaza Lunch Buffet -** (Chicken and beef fajitas with grilled red onions and bell peppers, Black beans (Vegetarian), Spanish rice (Vegetarian), Pork posole & calabacitas rancheras, Warm flour tortillas & butter, etc.) **End of meeting, enjoy lunch and Santa Fe**	**Sponsored by illumina**

05/27/2009 - Wednesday				
Time	**Type**	**Abstract #**	**Title**	**Speaker**
730 - 830am	Breakfast	x	**La Fonda Breakfast Buffet** (French "Texas Toast", Fluffy scramble eggs, Grilled breakfast potatoes, Applewood smoked bacon, breads, and fruit, etc.)	x
830 - 845	Intro	x	Welcome Intro	**Mary Neu**
x	Session Chair	x	Session Chair	Chair – Donna Muzny
845 - 930	**Keynote**	**FF0086**	**Fulfilling the Promise of a Sequenced Human Genome**	**Eric Green**
930 – 955	Speaker 1	FF0124	Defining Genome Project Standards in a New Era of Sequencing	**Patrick Chain**
955 – 1020	Speaker 2	FF0026	Human Microbiome Project Finishing	**Jessica Hostetler**
1020 – 1050	Break	x	Beverages provided	x
1050 – 1120	Speaker 3	FF0025	Applying Single Molecule Real Time DNA Sequencing	**Steve Turner**
1120 – 1150	Speaker 4	FF0146	Enabling True Biology with Helicos Single Molecule Sequencing	**Patrice Milos**
1150 – 120pm	Lunch	x	**Coronado Lunch Buffet** (Char-grilled chicken breast with barbecue-chipotle vinaigrette, Pan-seared rainbow trout fillet served w/ smoked yellow pepper coulis, Roasted garlic mashed potatoes & seasonal vegetables, etc.)	x
x	Session Chair	x	Session Chair	Chair – Johar Ali
120 – 140	Speaker 5	FF0038	Sequencing Complex Genomes Using Next Gen Platforms	**Tim Harkins**
140 – 200	Speaker 6	FF0093	SOLiD™ Finishing - High Throughput Sequencing, Assembly and Analysis	**Michael Rhodes**
200 – 220	Speaker 7	FF0132	Genomics and Expression	**Haley Fiske**
220 – 345	**Panel Discussion**	x	**Next Generation Sequencing Technology Panel Discussion**	Chair - Bob Fulton
345 – 400	Break	x	Beverages and snacks provided	x
400 - 500	**Tech Time Talks**	FF0013 FF0027 FF0033 FF0016	- A Software System for Validating and Orienting Sequence Contigs Using Optical Maps, - Sequencing & Finishing Pooled BAC Clones at WTSI with New Sequencing Technologies, - Whole-Genome Resequencing by Hybridization of Hemorrhagic Fever Viruses, - Relative Positioning of Scaffolds : a Challenge With New Sequencing Technologies	**Adam Briska,** **Darren Grafham,** **Sofi Ibrahim,** **Valerie Barbe**
500 – 700	**Posters - even #s**	x	**Poster Session**	x
600 - 800pm	**Meet & Greet Party**	x	**Meet & Greet Party - sponsored by Roche --- Food & Drinks**	**Sponsored by Roche**

Speaker Presentations (May 27th)

Abstracts are in order of presentation according to Agenda

FF0086

Keynote

Eric Green

Fulfilling the Promise of a Sequenced Human Genome

National Institutes of Health – National Human Genome Research Institute (NHGRI)

FF0124

Defining Genome Project Standards in a New Era of Sequencing

Patrick Chain and the International Genome Sequencing Standards Consortium

International Genome Sequencing Standards Consortium

For over a decade, genome projects, particularly whole microbial genome sequencing projects have adhered to a set of standards that can be generally relied on for purposes of sequence analysis by any interested third party. With the advent of novel and revolutionary sequencing technologies have come new applications of sequencing and a redefinition of traditional whole genome sequencing. However due to cheaper costs of generating near complete (draft) genomes, we are now faced with genome sequences in a wide variety of formats with varying levels of completeness and accuracy. We suggest a number of standard categories that these formats be placed in, such that the scientific community may understand the quality of, and thus fully benefit from, these projects.

FF0026

Human Microbiome Project Finishing

J. Hostetler1, M. FitzGerald2, R. Fulton3, D. Muzny4

1The J Craig Venter Institute, Rockville, MD, U.S.A.
2Broad Institute, Cambridge, MA, U.S.A
3The Genome Center at Washington University, St. Louis, MO, U.S.A.
4Baylor College of Medicine, Houston, TX, U.S.A

The Human Microbiome Project tackles the large goal of fully sequencing over 1000 reference genomes for diverse organisms living in and on the human body. Metagenomic samples representing various body sites from more than 50 healthy individuals will be compared to these reference genomes enabling us to better understand the nature of human and microbe interaction and its implications for health. The HMP calls for finishing 15% of the reference genomes while not defining the term "finished." The HMP Finishing Working group endeavored to set finishing levels that would be comparable across centers, independent of sequencing platforms and assembly software as well as tightly linked with the goals of the HMP Project.

Preliminary data have led us to evaluate an alternate approach where some finishing resources will be used to improve nearly all HMP reference genomes. Important, pioneer genomes would then be nominated for more thorough improvement as detailed below. We believe that this two-tiered approach is likely to provide the community with more data for about the same cost. The HMP finishing standards align closely with other major finishing groups but also include a higher level of detail in support of HMP's scientific goals. These provisional grades are based on a limited set of improved genomes and include: Improved-high-quality-draft, Annotation-grade, Non-contiguous finished, and Finished. Initial results show that Improved-high-quality-drafts have a 2.6 fold average increase in Contig N50 over High-quality-drafts for a very small amount of effort. We continue to accrue data for all the other defined levels in an effort to evaluate the utility of the grades in terms of the HMP goals and finishing on the whole. Here we present initial data generated by the project as well as a brief overview of the contributing centers' pipelines.

NOTES

FF0025

APPLYING SINGLE MOLECULE REAL TIME DNA SEQUENCING

Stephen W. Turner, PhD

Pacific Biosciences

SMRT (single molecule real time) DNA sequencing is a novel, high throughput method for sequencing DNA. It harnesses the intrinsic power of DNA polymerase enzymes as sequencing engines by eavesdropping on template-directed synthesis in real-time. Two critical technology components enable this process: The first is phospholinked nucleotides where, in contrast to other sequencing approaches, the fluorescent label is attached to the terminal phosphate rather than the base. The enzyme cleaves away the fluorophore as part of the incorporation process, leaving behind completely natural double-stranded DNA. The second critical component is zero-mode waveguide (ZMW) confinement technology that allows single-molecule detection at concentrations of labeled analogs relevant to the enzyme. Through the combination of these innovations, our technology allows the speed, processivity, efficiency and fidelity of the enzyme to be exploited. We show application of this technology to shotgun sequencing of human and bacterial DNA resulting in high consensus accuracy and unprecedented readlength. Because with Phospholink nucleotides the polymerase reverts completely to the initial state after each base sequenced the accuracy profile as a function of position within a read is flat. We will present a novel sample prep concept based on DNA hairpin ligation to double-stranded DNA that facilitates whole genome shotgun sequencing directly from genomic DNA with near-Poisson limited coverage uniformity and practically no GC bias. This sample prep will be demonstrated to enable consensus sequencing based on data extracted from just one molecule, allowing high accuracy sequencing at the molecular limit and without amplification. This capability is applied to detect SNPs in mixed samples with fractions as low as 1%. We will present a capability to perform paired-end reads from a single library with insert sizes that are adjustable at run-time, eliminating the need for multiple library creation for different insert sizes.

FF0146

Enabling True Biology with Helicos Single Molecule Sequencing

Patrice Milos and Helicos Colleagues

Helicos BioSciences Corporation, Cambridge, MA 02139

Helicos True Single Molecule Sequencing (tSMS)TM provides a unique view of genome biology through direct sequencing of cellular nucleic acids in an unbiased manner providing both quantitation and sequence information. Using a simple sample preparation involving no ligation or PCR amplification genomic DNA is sheared, tailed with poly A and hybridized to the flow cell surface containing oligo dT for initiating sequencing by synthesis. Helicos technology has been used successfully to sequence an array of bacteria representing the diverse genomic content of microorganisms, C. elegans and a human genome. The highly accurate quantitation of the platform is demonstrated in our gene expression, ChIP and copy number variation studies. The simple nature of the sample preparation has allowed the direct sequencing of nucleic acid from a variety of sample types including formalin-fixed paraffin embedded tissue and archival tissue samples. We have also optimized our sample preparation to allow preparation and sequencing from picogram quantities of nucleic acid – all important for maximizing researchers ability to perform important biological experiments with limiting biological sample amounts. Progress on the HelicosTM Genetic Analysis System performance including our genomic paired read data as well as details on our newest applications and methods and their use to study important biological samples will be discussed.

FF0038

Sequencing Complex Genomes Using Next Gen Platforms

Tim Harkins

Roche Diagnostics

Using 400 to 500 base pair sequencing reads and a combination of paired-end sequencing, it is now possible to generate draft assemblies of 100 to 500 megabases sized genomes. These assemblies can be generated in an automated fashion and often result in N50 contigs larger than 35 kb and N50 scaffolds in the megabase range. A series of strategies will be provided for sequencing complex genomes along with supporting examples. Additionally, technical updates to the Genome Sequencer FLX platform will be provided. Can 454 Sequencing extend beyond the 400 to 500 base pair shotgun sequencing reads to generate 800 to 1000 base pair reads?

FF0093

SOLiD™ Finishing - High Throughput Sequencing, Assembly and Analysis

Michael Rhodes

Applied Biosystems

The new SOLiD™ 3 System achieves new milestones in throughput in excess of 20 Gb of mate paired sequence data from a single run and 30-40 Gb of demonstrated throughput in Applied Biosystems R&D labs. Maintaining high accuracy, improvements in read length and unique mate-pair library strategies, the news system enables expanding applications from whole genome resequencing and SNP discovery to miRNA profiling. This presentation will review various applications that the new system capabilities enable, concentrating on those relevant to finishing including de novo assembly and whole genome resequencing for SNP and structural rearrangement discovery.

FF0132

Genomics and Expression

Haley Fiske

Illumina, Inc.

The Illumina Genome Analyzer, based on the Solexa massively parallel sequencing-by-synthesis technology, is being used for a broad set of functional genomics applications including chromosomal re-arrangements, to single nucleotide variations, variation in DNA methylation, whole transcriptome analysis, small RNA analysis, digital gene expression, DNA-protein, and DNA-RNA interaction analysis. Details on the current state of the technology as well as a summary of chromosomal resequencing studies, whole genome epigenetic changes, tissue-specific mRNA splice variatns and 5'-UTRs, microRNAs and DNA-protein interactions studies will be presented.

Panel Discussion Notes

Panel Discussion Notes

Panel Discussion Notes

FF0013

A Software System for Validating and Orienting Sequence Contigs Using Optical Maps

Adam Briska, Mihai Pop, Niranjan Nagarajan

OpGen, Inc., Gaithersburg, MD.

Although next generation sequencing methods are capable of rapidly producing high volumes of low-cost sequence data, the process of sequence finishing remains a significant challenge, due in part to the difficult task of determining the relative positions and orientations of sequence contigs, which is further complicated by the fact that sequence contigs can often contain mis-assemblies. Physical maps are powerful aids to resolving these complexities, but traditional BAC mapping can be expensive and time-consuming. Optical Mapping, on the other hand, is a technology which rapidly produces low-cost high-resolution ordered restriction maps of the entire genome independent of the sequence.

MapSolver is a graphical software tool which orients and validates sequence contigs against an Optical Map within minutes. MapSolver loads sequence information in from FASTA format and converts each sequence contig into an in silico Optical Map by finding the locations of restriction sites. Then, using a dynamic programming algorithm, it finds the optimal alignments of the in silico maps with the Optical Map. Based upon these alignments, MapSolver produces a visual representation of the relative orientations of these contigs and clearly shows any mis-assembled contigs. Additionally, this same information is easily exported into a tabular report.

This combination of the Optical Map data and MapSolver software can drastically reduce finishing times and costs.

FF0027

Sequencing and Finishing Pooled BAC clones at the WTSI with New Sequencing Technologies

Darren Grafham and Siobhan Whitehead

Wellcome Trust Sanger Institute, Hinxton, Cambridge, UK

The Wellcome Trust Sanger Institute (WTSI) is a world leader in genomic sequencing. Most notably it is responsible for the completion of one third of the sequence of the human genome, as well as the genomes of model organisms such as Mouse and Zebrafish and more than 90 pathogen genomes. The Zebrafish genome is due to be "essentially complete" by the end of 2009. The Illumina GAII and 454-FLX provide an opportunity to expediate this goal both efficiently and economically. They may also open up new avenues for future sequencing projects using a clone based approach.

Work is currently underway at the WTSI to test de novo sequencing, assembly and finishing of pooled Zebrafish BACs using the new sequencing technologies (NST). An initial test on 12 pooled BACs sequenced in 1 lane of an Illumina GAII, and separately on 454-FLX, has shown that although clone sequencing is possible, the current pipeline, assembly tools, and viewing tools need modification in order to cope with and manipulate the large amounts of data. Subsequent tests have stretched the tools to their limits and helped define the requirements for new tools, which will be presented.

The requirement of Phase III finished sequence coupled with the absence of templates to work with, leaves PCR or direct clone walks as the only options to progress from draft assembly to finished. The fragmented nature of new tech assemblies makes ordering PCR or direct clone walks both time consuming and in need of automation. Work is currently underway to implement the use of an automatic contig ordering and orientation tool (ABACAS) to generate PCR samples coupled with automatic PCR primer ordering to make a completely automated pipeline.

Presented here is a review of the trials carried out so far on pooled BAC clones, the development of the new visualisation and manipulation tools (Gap5 and Arcturus), and the development of an automatic PCR pipeline.

FF0033

Whole-Genome Resequencing by Hybridization of Hemorrhagic Fever Viruses

M. Sofi Ibrahim, M. Aitichou, J. Hardick, R. Abella

Virology Division, USAMRIID, Fort Detrick, Frederick, MD 21702

Periodic outbreaks of viral hemorrhagic fevers (VHF) have occurred in Africa over the past four decades with case fatality rates as high as 90%. The causative agents of these illnesses are classified as category A priority pathogens and pose a serious risk as bioterrorism agents. In this study, we designed whole-genome microarrays for rapid resequencing and identification of different species and strains of hemorrhagic fever viruses, including Ebola and Marburg viruses. We evaluated the array with the Ebola viruses Zaire, Sudan (Gulu) and Reston, and the Marburg viruses Angola, Ci67 (Popp) and Ravn. The results showed between 92.1% and 98.9% resequencing accuracy over the entire genomes of these viruses. The Basic Local Alignment Search Tool (NIH/NCBI-BLAST) returned best matching nucleotide sequence records from the NIH/NCBI GenBank sequence database compared with the microarray-generated sequences. These matching sequence records represented homologous species and closely related strains, with genome coverage between 95% and 100% and identity between 94% and 99%. Phylogenetic analysis placed all microarray-generated genome sequences into their correct taxonomic order among known members of Filoviridae. Thus, we demonstrated rapid, single sample, single assay whole-genome resequencing of HF filoviruses with high degree of accuracy. This approach could potentially be useful for detecting newly emerging viral species or strains.

FF0016

Relative Positioning of Scaffolds: a Challenge With New Sequencing Technologies

Barbe V., Mangenot S., Aury J .M., Castelli V., Chane-Woon-Ming B., Couloux A., Cruveiller S., Oztas S., Samson G., Vallenet D., Weissenbach J. and Wincker P.

CEA/DSV/IG/Genoscope, 2 rue Gaston Crémieux 91000 Evry, France

The progressive disappearance of Sanger sequencing methodology has induced new challenges essentially for reconstructing the correct genome organisation. Indeed, as genomic complexity and large repeat sequences very often result in a high supercontig number, large fragment sequence ends contribution allowed obtaining DNA sequence scaffolding. The Genoscope Finishing and (Bio)-Informatics laboratories have developed specific methodologies to improve chromosomal regions or microbial whole genome reconstructions.

To determine genomic regions, overlapping BACs are sequenced using multiplexing on 454GSflx sequencers (www.roche.com). After assembly using Newbler (www.roche.com) and/or Phrap (www.phrap.org) assemblers, BAC contigs organization is realized using automatic primer walks on BAC DNA. The consensus errors detection is performed using SNiPer, a program developed by staff members of the Laboratoire de Génomique Comparative (LGC, Genoscope), which remap the 454 reads on a reference molecule using the SSAHA2 aligner (www.sanger.ac.uk).

Microbial genome contig ordering is a more complex problem and finding a solution to it is essential for accurate comparative genomics analyses. At the present time, microbial genome sequences are obtained using a mix of 454 (single and paired-end for the assembly) and Solexa (for the polishing) technologies. The total number of scaffolds depends on large repeat sequence frequency in the genome and on the 454 paired-ends fragment size as well. We will describe here our efforts to perform genomic region sequencing and to improve complete genome assembly for microbial comparative genomics.

NOTES

NOTES

Poster Presentations (Even #s, May 27th)

FF0004

Cloning and Sequencing with Trace Amount of DNA on Roche/454 and Illumina platforms

Mansi Chovatia, Hope N. Tice, and Jan-Fang Cheng

US Department of Energy Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, California 94598 USA

As a user facility, The US Department of Energy's Joint Genome Institute, in collaboration with scientists around the world, are able to generate DNA sequences from a diversity of organisms and environmental communities. Often times, the amount of genomic DNA provided for library construction is very limited. It is imperative to develop a protocol to minimize the amount of genomic DNA required for library construction for the second-generation of sequencing platforms. We have begun constructing Roche/454 Shotgun and Illumina Paired End libraries with less than 1ug of genomic DNA by altering two key components from the standard operating protocol for library construction. The two key components that help minimize loss of genomic DNA are: shearing DNA via Covaris Adaptive Focused Acoustics™ (AFA) process instead of nebulization and utilizing Agencourt® AMPure® purification system to purify and select the size range of DNA fragments from contaminants and enzymes, with minimal loss of sample. This approach enables us to create Roche/454 libraries with as little as 300ng of genomic DNA and Illumina libraries with only 1ng of genomic DNA as the starting material.

FF0018

Improving Quality of Prokaryotic Genomes Using a Hybrid Sequencing Approach

Jean-Marc Aury1,2,3, Adriana Alberti1,2,3, Christophe Battail1,2,3, Corinne Cruaud1, Valérie Barbe1,2,3, Odile Rogier1,2,3, Sophie Mangenot1, Gaelle Samson1,2,3, Julie Poulain1, Véronique Anthouard1,2,3, Claude Scarpelli1,2,3, François Artiguenave1,2,3 and Patrick Winker1,2,3

1CEA, DSV, Institut de Génomique, Genoscope, 2 rue Gaston Crémieux, CP5706, 91057 Evry, France
2CNRS, UMR 8030, 2 rue Gaston Crémieux, CP5706, 91057 Evry, France
3Université d'Evry, 91057 Evry, France

Second generation sequencers are enabling the decoding of whole procaryotic genomes in less than a week at significantly lower cost than the classical Sanger method. However, each of these new sequencing technologies has been estimated to introduce new biases that yield a decrease in the quality of assemblies. Since these problems are method-specific, an appropriate mixture of sequences generated from different technologies may produce higher quality assemblies without requiring any Sanger-based input.

Another critical point for sequence assembly is the use of pair-mates information. Improvement of sample preparation of mate-paired DNA library may lead to an increase in assembly quality.

The efficiency of these different strategies to enhance genome assembly was assessed using the finished version of the gamma-protebacterium Acinetobacter baylyi. We estimated the impact on assembly quality of an optimised version of the protocol to prepare a mate-paired DNA fragment library and of the mixture of runs produced from Roche/454 GS-FLX and Illumina/Solexa Genome Analyzer.

FF0026

JCVI's Viral Finishing Pipeline: Integrating Automation and 454 to Increase Efficiency

E. Hine, K. Proudfoot, N. Fedorova, J. Sitz, D. Katzel, M. Kim, L. Overton, J. Hostetler, A. Djikeng, D. Spiro

The J. Craig Venter Institute, Rockville, MD, USA

The current viral finishing pipeline employs amplicon-based PCR Sanger sequencing, an assembly software suite called Elvira, then standard and custom primer design to close each genome. Each sample is thoroughly tracked using the Closure Task Manager (CTM), a tracking web interface. This pipeline has been used to sequence and close over 3300 Influenza A and B genomes from the beginning of 2005 to present. (http://msc.tigr.org/influenza/index.shtml). However, new automated programs are increasing the efficiency and the throughput of complete genome sequencing. Two new software programs that are currently in use are Plate Designer, which allows for easy loading of essential primer data into the database, and Vapor, which incorporates the existing Elvira assembly software, as well as new features that together allow it to load, assemble and validate Influenza samples.

Other programs, such as Contig Checker and Task Assigner are in testing and development. Contig Checker will examine samples for areas that need additional work and Task Assigner will assign the tasks needed to complete this added work. Our goal is to combine all these programs into a single complex suite for rapid and efficient complete genome sequencing of Influenza and other viruses with very limited to no manual interaction. This will increase efficiency to the point where only problem areas will require significant manual labor.

The automated pipeline is also being built and adapted to incorporate 454 reads, which are currently being tested for initial sequencing of viral genomes. The use of 454 reads for sequencing, Sanger reads for finishing, and the automated viral finishing pipeline throughout the process will significantly increase production, effectively decreasing costs and sample completion time.

FF0034

Whole Genome Sequencing for Rapid Identification of a New Bacterial Species

Andrew C. Stewart1, Brian Osborne2, Christopher P. Cook1, Amy Butani1, Shakia Dorsey1, Kristin M. Willner1, Kimberly Bishop-Lilly1, Timothy D. Read3 and Shanmuga Sozhamannan1

1 Genomics Department, Biological Defense Research Directorate, Naval Medical Research Center, Rockville, Maryland, United States
2 The BioTeam Inc., Middleton, Massachusetts, United States
3 Division of Infectious Diseases & Department of Human Genetics, Emory University School of Medicine, Atlanta, Georgia, United States

Rapid identification of known and unknown biodefense pathogens is essential for implementing countermeasures in a timely manner for saving lives. NexGen sequencing technologies have significantly reduced the cost of producing draft sequences rapidly. However, bioinformatic tools and pipelines for rapid analysis of the sequence data are still lacking.

DIYA (Do-It-Yourself Annotator) is a modular and configurable open source pipeline software, written in Perl, used for the rapid annotation of bacterial genome sequences. The software is currently used to take DNA contigs as input, either in the form of complete genomes or the result of shotgun sequencing, and produce an annotated sequence in GenBank file format as output. DIYA has been used at the NMRC Biological Defense Research Directorate for the annotation of 454 produced microbial sequences. The software is a component of Do It Yourself Genomics (DIYG), an open source consortium for genomic bioinformatics open source software. Here we show the utility of DIYA in annotation of draft genome sequences produced by 454 and compare it to a fully annotated, curated genome in NCBI db. We also demonstrate its utility in conjunction with MGAP (Mega Genomic Analysis Pipeline) for rapid phylogenetic assignment of a new/ unknown bacterial species.

FF0046

Efficient High Throughput Bacterial Assembly with Automated Plasmid Identification

Theresa Hepburn, Sarah Young, Aaron Berlin, David Heiman, Carsten Russ, Sakina Saif, Terrance Shea, Sean Sykes and Chad Nusbaum

Broad Institute of MIT and Harvard, 7 Cambridge Center, Cambridge, Massachusetts 02142, USA

With the explosion of next generation sequencing data and ambitious, multi-center research initiatives such as the Human Microbiome Project, efficient, scalable, and accurate assembly and analysis processes have become necessary. To address this we have implemented an automated bacterial assembly pipeline that reduces the human time requirement per genome assembly from four hours to less than 30 minutes. The pipeline inputs a standard set of read data and project information and outputs a completed *de novo* assembly, along with a comprehensive QC report of read quality, assembly metrics, contamination screen, and comparison to related genomes. In addition, we have built in automated analysis to identify potential plasmid sequences. Plasmids contain sequences of biological interest, such as antibiotic resistance genes, making rapid identification highly desirable. Plasmids differ from the chromosome(s) in characteristic ways. They are typically small, are often present at high copy number, frequently differ from the genome in GC content, and can contain known plasmid-associated genes. We evaluate assemblies for each of these characteristics. For short scaffolds (<300kb) we gather statistics on sequence coverage, GC composition, and results of alignment to the NCBI RefSeq plasmid database (ftp://ftp.ncbi.nih.gov/refseq/release/plasmid/); when available, related references are also aligned. The pipeline also checks for circularity in scaffold ends that overlap, or for reads or read pairs that suggest circularity. Finally, the pipeline generates a summary report with all metrics including calculated plasmid probability score for each scaffold. Our experience shows that alignment to related references and plasmid sequences along with read and read pairing based circularity evidence provide the strongest indicators of plasmid scaffolds. Using the full set of metrics we are able to make accurate predictions of plasmid scaffolds, as part of a process by which we can assemble and analyze hundreds of bacterial genomes per year.

FF0054

A Study of the Variation in 6 Diverse Strains of *Yersinia enterocolitica*

Danielle Walker and the Pathogen Sequencing Unit

Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SA, UK

Yersinia enterocolitica is a Gram-negative psychotropic bacterium, which causes acute gastro-enteritis and occasionally more serious disease in humans. Currently despite its importance as a global gastrointestinal pathogen only one genome sequence is available (Y. enterocolitica 8081), a highly pathogenic biotype 1B strain sequenced using Sanger capillary sequencing (Thomson et al. 2006). In addition to biotype 1B there are an additional 4 biotypes distinguished by distinct biochemical and virulence properties.

The selected isolates are from the remaining Yersinia enterocolitica biotypes, these in vitro biotypes group into 3 distinct grades of pathogen: a historically defined non-pathogenic group (biogroup 1A); a weakly pathogenic group unable to kill mice (biogroups 2 to 5), and a highly pathogenic, mouse lethal group (biogroup 1B). The murine yersiniosis infection model is very representative of human disease, and Yersinia represent an excellent genus to study the evolution of virulence.

This project aims to use a combination of sequencing technologies to study isolates from the remaining 4 Y. enterocolitica biotypes to support phenotypic studies of these strains. Data will be presented from these three platforms: 454/Roche GS20/FLX, 454/Roche Titanium FLX3 (PE) and the Illumina GAII Platform. Significant gap closure has been done using PCR largely generated using ABACAS (Algorithmic Based Automatic Contiguation of Assembled Shotgun Sequence), a new script currently being developed at the Wellcome Trust Sanger Institute.

Presented is the initial analysis across the 6 strains. This project will show how ‘finishing’ is developing and changing, how the expanding nature of projects matches the demands of the scientific community and how the approach to project management is changing to accommodate this.

FF0060

JGI Microbial Sequencing Process

Tijana Glavina del Rio1, Lynne Goodwin2, Susan Lucas1, David Bruce2, Alla Lapidus1, Nicole Shapiro1, Chris Daum1, Hope Tice1, Shweta Deshpande1

1. Joint Genome Institute, Production Genomic Facility, Walnut Creek, CA 94598
2. Joint Genome Institute, Los Alamos National Laboratory, Los Alamos, NM 87544

JGI Microbial process has undergone through major changes in the last year. The implementation of the new sequencing technologies, Roche 454 and Illumina, have been the key factor in migrating the microbial draft sequencing from the older Sanger pipeline to the new 454 and Illumina pipelines. This poster will present the current sequencing process for the microbial genomes at the PGF, Walnut Creek facility. Scope of Work for microbial projects will be discussed as well as the production workflow process steps from the DNA receipt to finishing. A brief summary of each process step will be provided as well as important statistics for the area. Scheduling will be addressed to show how the two separate pipelines are managed for project scheduling and synched in order to produce a complete dataset at the end. The poster will also provide information on the current number of microbes in process and the PMO forecast for the year.

This work was performed under the auspices of the US Department of Energy's Office of Science, Biological and Environmental Research Program, and by the University of California, Lawrence Berkeley National Laboratory under contract No. DE-AC02-05CH11231, Lawrence Livermore National Laboratory under Contract No. DE-AC52-07NA27344, and Los Alamos National Laboratory under contract No.DE-AC02-06NA25396.

FF0062

Next Generation Finishing Libraries at the Broad Institute

Anna Montmayeur, Chelsea Dunbar, Daniel Bessette, Harindra Arachchi, Amr Abouelleil, Margaret Priest, Niall Lennon, Lisa Zembek, Scott Anderson, Michael FitzGerald

The Broad Institute

The massively parallel sequencing capabilities that the next generation sequencing technologies provide have resulted in a greatly decreased cost per processed sample. However, this decreased cost is coupled with the requirement to also greatly increase sample input in order to fully take advantage of these savings. This challenge has necessitated the need to replace the way finishing samples at the Broad Institute are prepared in the laboratory from individual samples used by the Sanger-based capillary sequencers to the pooled sample libraries required by the next-gen sequencers. Here we evaluate two distinct sample library preparation methods used to generate finishing data using the 454 GS FLX Standard system. The first method relies on the addition of barcoded adaptors to unique PCR amplicons prior to pooling to create a library of distinguishable products with known sequence at both ends. The second method employs the creation of a randomly-sized amplicon library for scaffold gaps with known sequence only at one end using an anchored PCR approach. The methods and results of both approaches will be presented here.

FF0070

Gap Resolution: A Software Package for Improving Newbler Genome Assemblies

Stephan Trong, Kurt LaButti, Brian Foster, Cliff Han, Tom Brettin, and Alla Lapidus

DOE Joint Genome Institute, Walnut Creek, CA 94598

With the continued improvements of next generation sequencing technologies and their advantages over traditional Sanger sequencing, the Joint Genome Institute (JGI) has modified its sequencing pipeline to take advantage of the benefits of such technologies. Currently, standard 454 Titanium, paired end 454 Titanium, and Illumina GAII data are generated for all microbial projects and then assembled using the Newbler genome assembler. This allows us to efficiently produce high quality draft assemblies at a much greater throughput than before. However, it also presents us with new challenges. In addition to the increased throughput, we also have to deal with a larger number of gaps in the Newbler assemblies. Gaps in these assemblies are usually caused by repeats (Newbler collapses repeat copies into individual contigs, thus creating gaps), strong secondary structures, and artifacts of the PCR process (specific to 454 paired end libraries). Some gaps in draft assemblies can be resolved merely by adding back the collapsed data from repeats. To expedite gap closure and assembly improvement on large numbers of these assemblies, we developed a software package, called Gap Resolution, to perform the following process automation.

1. Identify and distribute the data for each gap into sub-projects.
2. Assemble the data associated with each sub-project using a secondary assembler, such as PGA.
3. Determine if any gaps are closed after reassembly, and either design fakes (consensus of closed gap) for those that closed or lab experiments for those that require additional data.

This software package was designed specifically to help automate the process of gap closure and assembly improvement in next generation assemblies. Use of this software on microbial genome assemblies has significantly alleviated manual finishing on each project. We are currently testing the software for use on fungal projects.

This work was performed under the auspices of the US Department of Energy's Office of Science, Biological and Environmental Research Program, and by the University of California, Lawrence Berkeley National Laboratory under contract No. DE-AC02-05CH11231, Lawrence Livermore National Laboratory under Contract No. DE-AC52-07NA27344, and Los Alamos National Laboratory under contract No.DE-AC02-06NA25396.

FF0074

SOLiD Next-Generation Genomics Services at SeqWright

Clayton M. Morrison, Xin-Xing Tan and Fei Lu

SeqWright Inc. 2575 West Bellfort, Suite 2001, Houston, TX 77054, USA

Successful commercial sequencing hinges on the ability of a service company to provide cost-effective outsourcing solutions to meet the needs of its customers. To be effective, service providers must accurately assess the commercial potential of new sequencing technologies. For providers of Next-Generation (Next-Gen) sequencing services, this entails validating a given platform and determining its applicability to specific customer requirements. Additionally, the provider must also build an extensive IT infrastructure to store and process the massive amount of sequence information generated by these platforms.

SeqWright was selected to be a preferred provider for the SOLiD Next-Gen sequencing platform from Applied Biosystems. SOLiD is a short-read sequencing instrument capable of generating staggering amounts of sequencing information. Indeed, the latest SOLiD upgrade allows the system to generate over 10 gigabases of mappable sequence per run for fragment libraries and over 20 gigabases of mappable sequence per run for mate-paired libraries. The capacity to generate these large datasets has proven to be particularly useful for SeqWright clients interested in whole genome resequencing as well as whole transcriptome analysis. For projects requiring less data per sample, sequencing slides can be subdivided into four or eight separate regions. With the inclusion of sample multiplexing, up to 320 individual samples can be analyzed in a single machine run. Such scalability allows SeqWright to meet the specific needs of customers for several additional Next-Gen sequencing applications including small RNA analysis, ChIP sequencing, targeted resequencing and SNP detection.

In addition to providing Next-Gen project consultation, library preparation and sequencing, SeqWright has also constructed an extensive IT infrastructure to analyze and store the large SOLiD datasets, which can exceed 4000 gigabytes of data per run. This allows SeqWright to assist clients with Next-Gen sequencing from project design through final data analysis.

FF0076

Finishing of New Technology Only Microbes and Fungi

Alicia Clum (aclum@lbl.gov), Hui Sun, Kurt LaButti, Brian Foster, Steve Lowry, Stephan Trong, and Alla Lapidus

Lawrence Berkeley National Lab

With the onset of new technology JGI has shifted its scope of work for microbes to 454 standard titanium, 454 paired end titanium, and illumina data for gap closer and quality improvement (polishing). Raw reads are assembled by Newbler to create a draft assembly that will be passed to finishers. An in-house developed software tool creates subprojects for each gap. In-silico attempts are made to close gaps using existing unassembled pyrosequence and Illumina data. Any remaining gaps are tackled by PCR based methods. These include standard PCR, bubble PCR, multiplex PCR, combinatorial PCR, and long range PCR. Once products are generated they can be sequenced, cloned or shattered as needed. Currently gap closing data is still generated using sanger. Eventually these gaps may be pooled and sequenced using illumina or 454. Once a genome is closed, illumina data is used to polish the genome. Any areas that are still substandard are subjects for resequencing. We extend this approach to fungal genomes.

This work was performed under the auspices of the US Department of Energy's Office of Science, Biological and Environmental Research Program, and by the University of California, Lawrence Berkeley National Laboratory under contract No. DE-AC02-05CH11231, Lawrence Livermore National Laboratory under Contract No. DE-AC52-07NA27344, and Los Alamos National Laboratory under contract No. DE-AC02-06NA25396.

FF0082

Can Pooled BACs be Finished Using 454 Sequencing Technology?

Sarah Pelan

The Wellcome Trust Sanger Institute

The Wellcome Trust Sanger Institute (WTSI) is a world leader in genome sequencing. To date it has been responsible for completing one third of the human genome sequence, and involved with sequencing over 100 pathogens and genomes of model organisms such as Mouse and Zebrafish. These genomes have been finished using standard Big Dye terminator sequencing, ABI capillary sequencers, assembly by Phrap (Green) and visualisation using Gap4 (Staden). With the Zebrafish genome due to be "essentially complete" by the end of 2009, it is hoped that 454 sequencing technology will help to achieve this goal both efficiently and economically.

A trial has been carried out on a pool of 12 BACs using 454-FLX technology. This initial pool contained Zebrafish and Pig BACs which had previously been finished using traditional Sanger sequencing methods. After 454 sequencing, the data was converted into a readable form in Gap4 using 454toGap (Bonfield). The aims of the trial were to look at the data de novo to assess whether normalisation of clones was needed prior to pooling, if there was any bias across certain regions, to establish each clone's contig number, to determine how many PCRs and Direct Walks were required to finish the clones, and to compare the sequence to Phase III finished data already available. Initially this work has been carried out manually, but trials are underway to investigate whether automated contig ordering, primer and PCR ordering tools can be used prior to clones reaching the finisher. A further trial using 454 titanium technology is underway, using unfinished BACs and fosmids. This pool includes clones which contain sequences that have previously proven difficult to finish.

Presented here are the findings of the 12 pooled BAC trial using 454-FLX technology at the WTSI, with a view to discussing how this impacts on clone based finishing at the WTSI in the future.

FF0096

Next Generation DNA Polymerases and Reverse Transcriptases

David Mead, Michael Nelson, Vinay Dhodda, Nick Hermersmann, Krishne Gowda, Darby Renneckar, and Tom Schoenfeld

Lucigen Corp., Middleton WI 53562

DNA polymerases (DNAP) are widely used for nucleic acid amplification, detection and sequencing. Reverse transcriptases (RT) are required to copy RNA into DNA. The most commonly used enzyme for Sanger sequencing is derived from Thermus aquaticus, whereas Bacillus sterothermophilus DNAP is used in the 454/Roche pyrosequencing platform. Retroviral RTs from AMV and M-MLV are limited by slow elongation rates, low processivity, poor fidelity, extensive strand switching, and low thermostability. Second and third generation instruments for massively parallel DNA sequencing can deliver megabases of data for a few dollars, with the promise of a human genome for a few thousand dollars in the near future. The development of a DNAP to match the technical capabilities of new instrument platforms has not kept pace. Achieving long and accurate reads using new solid phase template extension methods, terminator chemistries, and microfluidic flow technologies places new demands on the currently used enzymes. Polymerases with increased template affinity for DNA or RNA could provide important improvements in sequencing, amplification, and reverse transcription. We have engineered bacterial, archaeal, and viral polymerases to improve binding affinity. These derivatives show improved biochemical attributes in sequencing through regions of secondary structure and from single colonies or liquid cultures without prior purification.

Another promising area of polymerase development is the screening of metagenomic phage libraries for novel DNA- and RNA-copying enzymes. Phage polymerases are true replicases, unlike the microbial repair enzymes in common use. Replicases often possess biochemical properties favorable to in vitro applications, such as improved processivity and fidelity. Through homology and functional screens we have identified numerous viral replicases and some associated accessory proteins (e.g. helicases and primases), which are currently being developed as reagents that show promise to improve speed, throughput, accuracy, and reliability of DNA and RNA analysis. These DNA polymerases (and reverse transcriptases) from thermal aquifers are highly divergent from known enzymes and some are completely unique classes of enzymes, with little or no homology to Taq or Pfu DNAP. Several have been expressed to produce enzymes with novel properties and unique utility. The most extensively studied replicase allows high-fidelity, high efficiency PCR amplification of otherwise refractory sequences. An inherent reverse transcriptase activity allows single-tube, single-enzyme RT-PCR detection of RNA. This enzyme also allows Sanger sequencing of templates refractory to analysis using existing enzymes. A new class of DNA polymerases and affinity modifications promises to improve a variety of next generation sequencing and amplification applications.

FF0098

Macaque and Bovine Y Chromosome Finishing

Yan Ding1, Shannon Dugan-Rocha1, Christian J. Buhay1, Ziad Khan1, Michael E. Holder1, Qiaoyan Wang1, Wen Liu1, Jennifer Hughes2, Helen Skaletsky2, Donna Villasana1, Lynne Nazereth1, David Page2, Donna M. Muzny1 and Richard A. Gibbs1

1 Baylor College of Medicine, Human Genome Sequencing Center, Houston, TX 77030
2 Howard Hughes Medical Institute, Whitehead Institute, Massachusetts Institute of Technology, Cambridge, MA 02142

The BCM-HGSC finishing group has begun sequencing and finishing select regions on the Macaque Y and Bovine Y chromosomes. The Y chromosome is well known for its highly repetitive and largely duplicated regions. To deal with these challenges, traditional Sanger sequencing with a large insert library of 5-8kb has been employed. To date, approximately 373 bovine BACs and 60 macaque BACs have been prepped, sequenced and assembled in the BCM-HGSC pipeline. Of these, 291 bovine BACs and 43 macaque BACs totaling 60Mb of unique sequence have been completed at the highest quality or “gold standard”.

In order to accomplish the goal, initial draft assemblies with 8-10X Sanger coverage were generated with phredPhrap. At an average insert size of 175kb, these difficult BACs usually assemble into fewer than 10 contigs that may be misassembled over large duplicated regions. With the help of the larger insert read pairs, most of the duplicated regions can be manually sorted through tedious tearing and re-joining of contigs. Results have shown that the difficulty of sorting and binning many of these larger duplications is highly dependent on the number of copies present in the BAC. Closure of all remaining gaps and low quality regions is largely dependent on direct sequencing of BAC DNA which utilizes the whole genome amplified products from using the GE TempliPhi Sequence Resolver Kit. Although results may vary depending on the size and complexity of the target sequence, results have shown success rates of 95%-98% with an average Phred20 of 550bp. Transposon bombing or specialized shotgun libraries have been applied for difficult regions such as larger tandem repeats or GC rich regions. Finally, each completed BAC is validated by at least two restriction digestions to confirm assembly contiguity and accuracy. Additional direct BAC sequencing data and results of these strategies will be presented.

FF0100

Shotgun Assembly of a Repetitive Plant Genome

Jason Miller1, Sergey Koren1, Brian Walenz1, Granger Sutton1,
James Knight2, Thomas Jarvie2, Chinnappa Kodira2, Jason Affourtit2, Tim Harkins2

1 The J. Craig Venter Institute, Rockville MD USA
2 454 Life Sciences, Brandon CT USA

Plant genomes are notoriously repetitive. Genomic repeats prevent accurate reconstruction by the whole-genome shotgun assembly method. Repeats are particularly troublesome for next-generation sequencing (NGS) approaches. Since they generate reads shorter than Sanger reads, NGS sequencing offers less power to resolve repeats. We report the NGS sequencing and shotgun assembly of a repetitive plant genome. The cucumber (Cucumis sativus) Gy14 line is an inbred food crop. Its genome had been predicted to be diploid in 7 chromosomes spanning ~367Mbp. Approximately half the genome was thought to be composed of heterochromatic repeats. The genome was sequenced at 454 Life Sciences on the GS FLX Titanium platform with the XLR70 kit. Sequencing yielded 24.76M unpaired reads plus 11.44M paired-end reads from 3Kbp and 20Kbp libraries. Concordant assemblies were generated independently by two software applications, Newbler and Celera Assembler. Sequence alignment put 94% of both assemblies in one-to-one ungapped alignments and 99% in gapped alignments. The assemblers each generated about 200Mbp of consensus scaffold sequence. Of 427 available cucumber mRNA sequences, 98% mapped to the assembly at a 90% identity threshold. Both assemblers generated, in addition, about 150Mbp of mini-assemblies unassigned to scaffolds. Analysis of K-mer content revealed the unassigned sequence is repetitive and dissimilar from the scaffold population. A majority of the unassigned sequence maps to a small span of the scaffolds. We conclude that some scaffolds enter long tracts of genomic repeat that are otherwise left unassembled. Despite the repetitive nature of this genome, our methods segregated the heterochromatin and resolved the euchromatic sequence.

FF0106

The HMP Microbial Assembly Pipeline at The Genome Center – Washington University School of Medicine

Chad M. Tomlinson , Wesley C. Warren, Patrick Minx, Bob Fulton, Lucinda Fulton, Erica Sodergren, George Weinstock, Elaine R. Mardis, and Richard K. Wilson1

The Genome Center at Washington University School of Medicine, St. Louis, Missouri

The Genome Center at Washington University School of Medicine has played a major role in the sequencing, assembly, and annotation of genomes associated with the Human Gut Microbiome Initiative (HGMI) and the Jumpstart portion of the Human Microbiome Project (HMP) . Since the initiation of the HGMI project in 2006, there have been many adjustments to our microbial assembly pipeline at the Genome Center. These adjustments correlate to improvements in next generation sequencing technology as well as assembler innovation. Initially, we generated Newbler assemblies of 454 fragment data and then used the PCAP assembler to combine Sanger data with the fasta sequences from the Newbler assembly. The advent of 454 paired-end technology as well as the longer read length Titanium fragment technology, has enabled us to generate de novo Newbler assemblies, which meet or exceed the high quality draft standards established by the HMP consortium. The assemblies are screened for extra-divisional and host contaminants using BLAST and evaluated based on established minimum contiguity and coverage metrics. The assemblies are then sent to our manual sequence improvement group for evaluation. This involves the breaking apart of misassembles, joining contigs, and the order and orientation of the remaining contigs. We submit both a draft and improved assembly to GenBank and submit gene annotation for the improved assembly. We are presently investigating the potential transition of our microbial pipeline to a de novo Illumina assembly generated with the Velvet assembler. This would represent a tremendous cost savings over our current 454 de novo platform.

FF0152

Metagenomic studies of thermophilic and psychrophilic microbial communities

Cristina Takacs-Vesbach

Department of Biology, University of New Mexico, Albuquerque, NM 87104

Our understanding of the diversity and distribution of microorganisms in the natural environment has been radically changed by the application of molecular biology techniques to microbial ecology. We are presently on the verge of taking another major leap in our understanding of microbial ecology because of environmental metagenomics. We are using metagenomics to investigate the diversity and function of microbial communities from extreme environments. Our first project is focused on a hydrothermal spring named Bechler, located in the SW quadrant of Yellowstone National Park that contains a novel, but very simple thermophilic community. We will combine 40 Mb of Sanger sequence with 100 Mb of pyrosequencing data to determine the phylogeny, metabolic potential, and in situ role of microorganisms in this community. Additionally, we will be conducting a comparative metagenomic analysis of communities from wetted and dry soils of the McMurdo Dry Valleys of Antarctica where a distinct endemic community of pycrorphiles exists. Additionally, complementary cultivation and microscopy approaches will be employed. Ultimately, we hope to use these data to develop complementary tools such as expression microarrays and proteomic studies to broaden our ability to construct meaningful in situ experiments.

FF0158

Sequence Enrichment using Droplet-based PCR

Keith Brown and Take Ogawa

Rain Dance Technologies Inc.

In order to maximize the efficiency of the 2nd Generation of DNA sequencers, a strategy for enriching biologically relevant loci on the mega base scale has been developed by RainDance Technologies. RainDance uses micro-droplet PCR to selectively amplify thousands of loci simultaneously in a single PCR tube. This approach produces a uniform representation of targeted loci and an unbiased representation of alleles. Here we present an overview of the application and sequencing results targeting a list of exons involved in the core signaling pathway of pancreatic cancer.

Poster Session Notes

Poster Session Notes

Meet and Greet Party

600pm – 800pm, May 27th

Sponsored by Roche Diagnostics

Enjoy!!!

05/28/2009 - Thursday				
Time	**Type**	**Abstract #**	**Title**	**Speaker**
730 - 830am	Breakfast	x	**Santa Fe Breakfast Buffet** (Scrambled eggs with a choice of three accompaniments on the side -chilaquiles with green chile and cheese, chorizo sausage and roasted green chile, Grilled breakfast potatoes, applewood-smoked bacon and warm flour tortillas, assorted breads and fruits, etc.)	x
830 - 845	Intro	X / FF0156	Welcome Back + **X prize announcement**	**Chris Detter / Larry Kedes**
x	Session Chair	x	Session Chair	Chair – Mike Fitzgerald
845 - 930	**Keynote**	**FF0087**	**Sequencing Complex Regions of the Genome: Disease & Evolutionary Impact**	**Evan Eichler**
930 - 955	Speaker 1	FF0125	Tools for Managing and Comparing Assemblies	**Deanna Church**
955 - 1020	Break	x	**Beverages provided**	x
1020 -1045	Speaker 2	FF0140	Assembly and Finishing of Small and Large Genomes	**Jim Knight**
1045 -1110	Speaker 3	FF0047	ALLPATHS: Assembling Large Genomes with Short illumina Reads	**Sante Gnerre**
1110 - 1135	Speaker 4	FF0130	Adapting Celera Assembler to 454 Platforms and Automated Finishing	**Sergey Koren**
1135 - 1200	Speaker 5	FF0090	AutoSeq: Auto-Assembly Pipeline in a Small DNA Sequencing Core Facility	**Jan Kieleczawa**
1200 - 130pm	Lunch	x	**New Mexican Lunch Buffet** (Pork tenderloin achiote-rubbed and char-grilled with tomatillo-chipotle sauce, your choice of either Chicken or Cheese enchiladas with red or green chile, etc.)	x
x	Session Chair	x	Session Chair	Chair – Patrick Chain
130 -150	Speaker 6	FF0024	Towards the $1000 Genome - the Impact of New DNA Technologies on Finishing	**Niranjan Nagarajan**
150 -210	Speaker 7	FF0035	To 'Finish' or Not to 'Finish'-the $64K Question in Genomics - Data Mining Using 454 Draft Sequences	**Shanmuga Sozhamannan**
210 – 230	Speaker 8	FF0066	New Technology Drafts: Production and Improvements	**Alla Lapidus**
230 – 250	Speaker 9	FF0097	Microbial Finishing at BCM-HGSC	**Shannon Dugan**
250 – 310	Speaker 10	FF0081	Prescreening of Data from GAii Sequencer Resulting in High-Quality Results in Genome Finishing	**Cliff Han**
310 – 330	Break	x	**Beverages and snacks provided**	x
330 – 430	**Tech Time Talks**	FF0110 FF0104 FF0096 FF0071	- Throughput to Insight... and Tackling the Messy Bits Between, - Sequence Enrichment Applying RainDance Technology, - Next Generation DNA Polymerases and Reverse Transcriptases, - Optimizing DNA Shearing for Next-generation Sequencing	**Jarret Glasscock,** **Vincent Magrini,** **David Mead,** **James Laugharn**
430 - 630	**Wine & Cheese**	x	**Beverages, Wine & Cheese provided - sponsored by OpGen**	x
430 - 630	**Posters - odd #s**	x	**Poster Session with Wine & Cheese - sponsored by OpGen**	x
630 - bedtime	on your own	x	Dinner and night on your own - enjoy	x

Speaker Presentations (May 28th)

Abstracts are in order of presentation according to Agenda

FF0156

The $10M Archon Genomic X PRIZE and Data Validation

Larry Kedes (1,2) and Granger Sutton (3)

(1) X PRIZE Foundation, Playa Vista, CA
(2) Institute for Genetic Medicine, University of Southern California, Los Angeles, CA
(3) J. Craig Venter Institute. Rockville, MD

The purpose of the Archon Genomic X PRIZE competition is to encourage the development of radically new technology that will dramatically reduce the time and cost of sequencing genomes, and accelerate a new era of predictive and personalized medicine. Through the competition the X PRIZE Foundation (XPF) aims to enable the development of low-cost diagnostic sequencing of human genomes.

The $10 million X PRIZE for Genomics prize purse will be awarded to the first Team that can develop a method and use it to sequence 100 human genomes within 10 days or less at a cost of $10,000.00. There are three primary evaluation criteria: 98% completeness compared to a reference, 99.999% accuracy (1 error per 100 kbp), and production of a full diploid genome (2 complete copies of each chromosome except in the case of a male sample where single copies of the X and Y chromosomes are to be provided). The XPF intends to develop a robust and routine approach to evaluating the quality, accuracy and completeness of producing 100 human genome assemblies .

We will outline approaches to a validation methodology that involves carefully choosing DNA samples and validation methods that will test the contestants' capability to sequence and assemble genomes in a consistent, accurate, and unbiased manner. The validation methods are constrained by the assumption that validation costs should be a small fraction of the X PRIZE award.

NOTES

FF0087

Keynote

Evan Eichler

Sequencing Complex Regions of the Genome: Disease & Evolutionary Impact

Department of Genome Sciences, University of Washington

FF0125

Tools for Managing and Comparing Assemblies

Deanna M. Church, Nathan Bouk, Cliff Clausen, Victor Sapojnikov, Boris Fedorov, Yuri Kapustin, Josh Cherry, Wratko Hlavina, Charlie Xiang, Alexey Sidelnikov, Alex Astashyn, Olga Ermolaeva, Donna Maglott, Mike DiCuccio, Paul Kitts and Avi Kimchi

National Center for Biotechnology Information, Bethesda, MD

Genome assemblies have not been formally tracked or accessioned as a unit at NCBI. Typically tracking has been managed via names and date stamps that were consistent within a given browser or FTP site. As part of the redesign of our genome management and pipeline system, we have constructed a new system to manage submission and tracking of genome assemblies. The new system allows for simplified processing of AGP files as well as the assignment of unique identifiers by which assemblies can be tracked. Submitter defined metadata, such as names, methods, etc, is associated with the assembly and can be submitted such that all clients of the data will have the exact same information.

In addition to systems for tracking assembly data, we have also developed several tools that allow us to compare assemblies, currently either unrelated assemblies from the same organism (such as the human reference Build 36 vs. the Venter assembly) or different versions of the same assembly (such as the human reference in Build 35 vs. Build 36). The basis of these comparisons is assembly to assembly alignments. The alignments are generated first by BLAST. The BLAST hits are analyzed by a greedy algorithm that reconciles alignments across multiple sequences in order to optimize scoring and coverage. Typically, two sets of alignments are generated, a one to one set where any given sequence in one assembly has at most one alignment in the other assembly and a many to many alignment set that allows duplicated sequences to be included in the alignment set. This is useful for identifying sequences that have collapsed or expanded between two assemblies. Additionally, these alignments serve as the basis for an annotation remapping service we have recently developed. The alignments are also used to manage genome annotation, and differences in transcript placement can also be used to measure differences between assemblies. Additionally, these tools allow us to produce consistent annotation across independent assemblies. NCBI Human Build 36.3 was the first to use assembly alignments to provide this consistent annotation (across the reference, Celera and HuRef assemblies). Together, this suite of tools allows for robust tracking of assemblies as well as rapid identification of assembly differences.

FF0140

Assembly and Finishing of Small and Large Genomes

Jim Knight

Roche Diagnostics

Recent advances in GS FLX sequencing have led to the reality of sequencing 4 bacterial genomes in a single run and gigabase eukaryotic genomes in a few weeks, with enough information content to generate high quality de novo assembly results from the data. This talk will describe the advances and improvements to the Newbler assembler to try to keep pace with the rate of sequencing, and to generate 1 scaffold bacterial assemblies and publication-equivalent large genome assemblies. Some of the topics included are the generation of assembly results that fit into existing finishing pipelines, the parallelization of the assembler, the scale up to handling gigabase genome datasets and visualization tools to help the in-silico closing of gaps.

FF0047

ALLPATHS: ASSEMBLING LARGE GENOMES WITH SHORT ILLUMINA READS

Sante Gnerre, Iain MacCallum, Dariusz Przybylski, Joshua Burton, Filipe Ribeiro, Genome Sequencing Platform, Chad Nusbaum, and David B. Jaffe

Broad Institute of MIT and Harvard, Cambridge, MA

New short-read sequencing technologies offer massive economic savings and throughput increases: mammalian genomes assembled from Sanger-chemistry reads early in this decade, for example, cost roughly $50 million each, while today the equivalent genomic coverage from massively parallel sequencing technologies would cost less than $50,000.

Accordingly, we aim to develop the ability to generate high quality assemblies of large genomes using short reads, an extremely challenging problem that remains unsolved.

Our test cases for large genome assembly from short reads is the 450 Mb genome of Gasterosteus aculeatus, the three-spined stickleback fish, which we had previously sequenced and assembled using traditional Sanger-chemistry data.

The primary data for this project consist of paired 100 base reads from two libraries, sequenced with the Illumina platform: a 180 base fragment library, and a 4000 base 'jumping' library. Because the two reads in a fragment-library pair almost always overlap, we can join them into a single 'super-read' with the aid of a corroborating third read. This raises the effective read length to 180 bases. We then proceed to assemble the 'super-reads' into an initial large De Bruijn graph, which we use a starting point for the localization process. After localization, we run in parallel thousands of local assemblies, which are merged in the last step of the assembly algorithm.

Here we present the initial draft of the stickleback assembly. We employ an enhanced version of the ALLPATHS algorithm that we have used previously to assemble small genomes from 36 base reads; we have adapted it to work with the new long read Illumina data described above. We rigorously assess the quality of our assembly by comparing it to the preexisting Sanger-data assembly.

FF0130

Adapting Celera Assembler to 454 Platforms and Automated Finishing

Sergey Koren, Jason Miller, Eli Venter, Brian Walenz, and Granger Sutton

The J. Craig Venter Institute, 9712 Medical Center Drive, Rockville MD 20850

Automated finishing can benefit from the mature and feature-rich software that exists for de novo assembly. The Celera Assembler, a de novo whole-genome shotgun assembly package, has enabled publications of microbial and mammalian genomes for 10 years. We enhanced the software to support the 454 FLX and Titanium platforms. We present our experience applying Titanium data to de novo whole-genome assembly. We have also adapted the software to support automated finishing. A new optional feedback loop enlarges contigs without additional sequencing. After an initial run, a new module identifies sequence that was presumed repetitive due to high coverage. For sequences that assemble to a single locus, the feedback loop boosts confidence in their uniqueness. A second run of the assembler exploits this information. This technique boosted contig and scaffold size, as well as mate constraint satisfaction, on a large and repetitive plant genome. Another new feature exploits co-location constraints placed on finishing reads. It was tested at JCVI, where a pipeline called AutoClosure generates PCR primers around gaps and then generates finishing reads from the PCR products. The revised Celera Assembler operates on the shotgun and finishing reads together, cognizant that reads from each PCR product should assemble between their primer sites. Tested on bacterial genomes finished at JCVI, the software correctly placed the finishing reads and improved assembly metrics. Now, this powerful software for de novo assembly can be leveraged for automated finishing.

Funding:
The National Institutes of Health (NIH).

FF0090

AutoSeq: Auto-Assembly Pipeline in a Small DNA Sequencing Core Facility

Jan Kieleczawa1, Vladimir Kubatin2, Don Koffman3 and Tony Li1

1 Wyeth Research, Cambridge, MA;
2 InforSense, Cambridge, MA; 3 Informed Solutions, Brookline, MA

Current DNA sequencing effort at the core facility of Wyeth Research heavily relies on manual contig assembly and editing which is refered to as “analysis component” of our operation and typically 60-70% of analysts time is devoted to this task. To alleviate this bottleneck we have developed a new auto assembly pipeline, AutoSeq. It combines series of publicly available PERL scripts (PHRED/PHRAP/CrossMatch) and in-house developed bioinformatics tools. Assembled high quality contigs and rudimentary annotation analysis (e.g. “your consensus sequence 100% matches reference sequence”) are pipelined via CAF mechanism into our 4D LIMS DNA sequencing database and is available at any to requestors. If manual intervention is needed, the CAF is imported into Sequencher project and additional analysis performed. Our initial comparison indicates that at least 50% of sequencing projects can be completed with no, or minimal, human intervention and with time savings of 60-70% compared to manual operations.

NOTES

NOTES

FF0024

Towards the $1000 Genome - the Impact of New DNA Technologies on Finishing

Niranjan Nagarajan

University of Maryland

While new sequencing technologies have ushered in an era where microbial genomes can be easily sequenced, the goal of routinely producing high-quality draft and finished genomes in a cost-effective fashion has still remained elusive. A crucial impediment has been the lack of tools that can augment the fragmented assemblies from shotgun sequencing with additional sequencing and map based information. In this work, we describe our experience in successfully producing finished and nearly-finished assemblies for a range of microbial genomes. These genomes were obtained with surprisingly little investments in terms of time, computational effort and sequencing and finishing cost compared to other recent projects. In particular, we highlight the use of contig graph and optical map information for cost-effective scaffolding and finishing of genomes and we describe a set of easily accessible tools to exploit this data.

FF0035

To 'Finish' or Not to 'Finish'-the $64K Question in Genomics - Data Mining Using 454 Draft Sequences

Shanmuga Sozhamannan1*, Arya Akmal1, Maureen P. Kiley1, Shannon M. Lentz1, Nicole M. E. Nolan1, Kristin M. Willner1, Christopher P. Cook1, Amy Butani1, Shakia Dorsey1, Trupti N. Brahmbhatt1, Al Mateczun1, Timothy D Read3 and Kimberly A. Bishop-Lilly1

1 Genomics Department, Biological Defense Research Directorate, Naval Medical Research Center, Rockville, Maryland, United States
2 Division of Infectious Diseases & Department of Human Genetics, Emory University School of Medicine, Atlanta, Georgia, United States

NexGen sequencing technologies can produce high quality, high coverage genomic data at a much lower cost and with a smaller footprint than traditional high cost, labor-intensive Sanger sequencing. The question becomes how to extract useful information from genome sequences produced by NexGen technologies in clinical and biothreat situations where timely identification of genetic variations and deliberate genetic modifications are crucial in saving lives. We have sequenced bacterial strains with known and unknown genetic modifications using the 454 GS20, 454 FLX and 454 Titanium sequencing systems to determine if they can identify such variants. Using 454 FLX, we sequenced a □GerH mutant of a Bacillus anthracis Sterne strain generating 95Mb total and 84 large contigs. We observed a total of 100 differences from a reference genome, including indels. Of these, 32 are high confidence single nucleotide polymorphism (SNP) calls (meaning that the SNP is present in □85% of reads). 25 of these were independently confirmed by SOLiD sequencing as well and 4 were likely SOLiD false negatives.

While 'finishing' genomes may be necessary for definitive elucidation of genome content and structure, and for ultimate forensic purposes, so called 'draft' genome sequences produced by 454 may be sufficient to identify genetic changes such as indels and SNPs. We are also confirming these SNPs by Sanger sequencing. Given the 'finished' genome error rate of 1 in 100,000 (phred score of 50), we would have expected to observe 50 errors in a traditional Sanger sequenced genome with no gaps. By comparison, the 'draft' genome data reported here appears as accurate as traditional finished genome sequence. Our results suggest that high quality draft 454 genomic sequences can be used for identification and SNP detection without the added cost and labor involved in traditional genome finishing. We are currently assessing the performance of 454 Titanium sequencing for SNP detection in comparison with these other platforms and results will be discussed.

FF0066

New Technology Drafts: Production and Improvements

Alla Lapidus, Tom Brettin, Jan-Fang Cheng, Alicia Clum, Alex Copeland, Chris Daum, Chris Detter, Brian Foster, Tijana Glavina del Rio, Cliff Han, Kurt LaButti, Matt Nolan, Simon Roberts, Hui Sun, Stephan Trong, Susan Lucas

DOE-JGI, Production Genomics Facility, Walnut Creek, CA
DOE-JGI, Las Alamos National Laboratory, Los Alamos, NM

The Joint Genome Institute (JGI) is a world wide leader in microbial genome sequencing. Our microbial sequencing pipeline was rebuilt to accommodate the benefits of next generation sequencing platforms. By applying QC analysis and by using the Newbler assembler we are able to produce high quality reliable skeletons of genomes from a combination of standard and paired ended 454 data. Illumina data is used to help improve the overall consensus quality and to close secondary structure related gaps. Two in-house developed software tools – Polisher and Gap Resolution (see poster presented by K. LaButti and S. Trong) – are used to resolve repetitive gaps in Newbler assembly and to correct errors in the produced consensus. The genome finishing process relies on newly developed bubble PCR approach (bPCR; see poster, presented by H. Tice) and traditional lab techniques such as long-range and multiplex PCRs (see poster presented by A. Clum). The created pipeline has been applied to numerous sequencing projects over the course of last year. We are developing strategies to integrate new sequencing technologies in fungal genome assembly and finishing. The results and protocols of these strategies will be presented at a later date.

This work was performed under the auspices of the US Department of Energy's Office of Science, Biological and Environmental Research Program, and by the University of California, Lawrence Berkeley National Laboratory under contract No. DE-AC02-05CH11231, Lawrence Livermore National Laboratory under Contract No. DE-AC52-07NA27344, and Los Alamos National Laboratory under contract No. DE-AC02-06NA25396.

FF0097

Microbial Finishing at BCM-HGSC

Shannon Dugan-Rocha, Yan Ding, Christian J. Buhay, Michael E. Holder, Xiang Qin, Vandita Joshi, Joe Petrosino, Sarah Highlander, Donna Muzny and Richard Gibbs

Baylor College of Medicine, Human Genome Sequencing Center, Houston, TX, 77030

BCM-HGSC is one of four major centers involved in the Human Microbiome Project. To date, 72 microbial genomes have been assembled as part of the jumpstart portion of this project. These microbial genomes were assembled with the Roche/454 Newbler software, generating N50 contig lengths averaging 51kb and N50 scaffold lengths averaging greater than 700kb. Solexa/Illumina data was also incorporated in order to leverage the strengths of each NexGen technology for error correction and assembly contiguity. Of these assembled microbes, 24 have been completed at various finishing levels. These finishing levels range from automated and or/manual work that provides an “improved” quality over the initial draft sequence to high quality or “gold standard” finished genomes with no unresolved or unconfirmed regions.

For the purpose of finishing, we have employed all pipeline advances now developed for new technologies including new assembly/mapping methods (Newbler and Mosaik), pre-finishing (Autofinish) modifications, along with gap closure methods developed for NexGen assemblies. Additional tools to further automate the process of sorting and binning repeats by utilizing updated versions of the Newbler assembly software has significantly aided in removing ambiguity from contig ends. Editing program upgrades, including implementation of CONSED 19.0 are now in transition to the finishing pipeline for all finishing activities. This latest version of CONSED has the ability to incorporate 454 reads for viewing and editing in CONSED as well as the ability to incorporate Solexa reads and the functionality to navigate these deep coverage assemblies for sequence discrepancies.

Further work testing accuracy, effectiveness, and benefit of an additional finishing category, Annotation Directed/Validated Finishing has also been conducted. This level of finishing is a process in which possible sequence improvement targets are identified through automated gene annotation. Five genomes completed at BCM-HGSC were submitted to the Gene Prediction Improvement Pipeline (GenePRIMP) for analysis. These five genomes: Mobiluncus curtisii, Gardinerella vaginalis, Lactobacillus jensenii, Staphylococcus aureus MN8, and Streptococcus pneumonia 19A, were submitted to the pipeline at varying levels of improvement in order to obtain information on potential problem areas. These problems included premature stop codons, truncated genes and split genes. Flagged areas were examined and categorized according to those that could be confirmed by manual review and those that would need to be resolved by additional directed reactions. Efforts to fully define this category are ongoing pending further analysis in addition to developments to reliably target additional anomalies including possible frameshifts.

FF0081

Prescreening of Data from GAii Sequencer Resulting in High-Quality Results in Genome Finishing

Cliff S. Han1, Stephan Trong2, Kurt M. LaButt2, Brian Foster2, Olga Chertkov, Alla Lapidus2

1 DOE Joint Genome Institute, Bioscience Division, Los Alamos National Laboratory, Los Alamos, NM 87545.
2 JGI Production Genomics Facility, 2800 Mitchell Drive, Walnut Creek, CA 94598.

Joint Genome Institute from Department of Energy has completely sequenced more than 300 bacterial genomes, most of which are bioenergy, bioremediation, and carbon sequestration related microbes. With the next generation sequencing technologies emerging as cheap high throughput sequencing instruments, we modified our process to sequence bacterial genome as the following: 10-20 x coverage from the titanium 454 sequencer, 10 x coverage of paired ends from titanium, and 50 – 100 x from Illumina GAii sequencer. Data from 454 machine is assembled, and that from Illumina machine is used to polish potential errors from Newbler assembly. In the transition stage, we have sequenced some project together with Sanger technology as well. This allowed us to compare the quality of data from new technologies with the gold standard of data from ABI3730. We first used the data from Illumina directly to correct assembly from 454 data. Except the right corrections, this process introduced some new errors as well when using high quality Sanger sequencing as standard. We then prescreen the data from Illumina machine by removing low quality sequences, homo-nucleotide sequence artifacts, and reduce redundancy of identical reads that is over amplified. The screened data introduced very few errors in less polishing process, less than one in 5 Mb. Our experience tells that the data from new sequencing technologies need to be screened for better results in analyzing the alignment result. If not, miscall of SNPs could be increased. We also had experience that screening low quality data could result in better assembly with short reads assembler, such as Velvet.

This work was performed under the auspices of the US Department of Energy's Office of Science, Biological and Environmental Research Program, and by the University of California, Lawrence Berkeley National Laboratory under contract No. DE-AC02-05CH11231, Lawrence Livermore National Laboratory under Contract No. DE-AC52-07NA27344, and Los Alamos National Laboratory under contract No.DE-AC02-06NA25396.

NOTES

FF0110

Throughput to Insight... and Tackling the Messy Bits Between

Jarret Glasscock, Ryan Richt, Matt Hickenbotham

Cofactor Genomics, Saint Louis, MO USA

Short-read platforms have recently moved us into the 50-100bp range and this paired with higher densities have resulted in individual sequencing runs that are tens of Gb in size (sure to both be updated by the time this is read). These advances have allowed many researchers to move from applying these technologies from re-sequencing studies to larger and more complex de-novo genome, deep expression, and complex transcriptome characterizations.

Undoubtably, we have a lot of work to do in order to shorten the gap between these initial sequencing products and the quality, contiguity, and utility that researchers have come to expect after years of working with model organism assemblies and manually annotated gene sets. We present a number of sequencing projects, our progress and approaches to shortening this gap, and our work on developing analysis and visualization tools to keep pace.

FF0104

Sequence Enrichment Applying RainDance Technology

Vince Magrini, Ryan Demeter, Jon Armstrong, Todd Wylie, Rick Wilson, and Elaine Mardis.

Washington University in St. Louis

RainDance Technologies (RDT) provides a micro droplet-based platform presently designed toward PCR targeted amplification in nanoliter volumes. As part of the early access program, we've evaluated libraries (primer sets) targeting 384 RDT and in-house amplicons using both genomic and whole genome amplified input DNA. Here, we will report on the PCR representation and variant detection using Illumina sequence data. Additionally, RDT designed a library set targeting 4000 amplicons, and we'll report on target representation within in this data set.

FF0096

Next Generation DNA Polymerases and Reverse Transcriptases

David Mead, Michael Nelson, Vinay Dhodda, Nick Hermersmann, Krishne Gowda, Darby Renneckar, and Tom Schoenfeld

Lucigen Corp., Middleton WI 53562

DNA polymerases (DNAP) are widely used for nucleic acid amplification, detection and sequencing. Reverse transcriptases (RT) are required to copy RNA into DNA. The most commonly used enzyme for Sanger sequencing is derived from Thermus aquaticus, whereas Bacillus sterothermophilus DNAP is used in the 454/Roche pyrosequencing platform. Retroviral RTs from AMV and M-MLV are limited by slow elongation rates, low processivity, poor fidelity, extensive strand switching, and low thermostability. Second and third generation instruments for massively parallel DNA sequencing can deliver megabases of data for a few dollars, with the promise of a human genome for a few thousand dollars in the near future. The development of a DNAP to match the technical capabilities of new instrument platforms has not kept pace. Achieving long and accurate reads using new solid phase template extension methods, terminator chemistries, and microfluidic flow technologies places new demands on the currently used enzymes. Polymerases with increased template affinity for DNA or RNA could provide important improvements in sequencing, amplification, and reverse transcription. We have engineered bacterial, archaeal, and viral polymerases to improve binding affinity. These derivatives show improved biochemical attributes in sequencing through regions of secondary structure and from single colonies or liquid cultures without prior purification.

Another promising area of polymerase development is the screening of metagenomic phage libraries for novel DNA- and RNA-copying enzymes. Phage polymerases are true replicases, unlike the microbial repair enzymes in common use. Replicases often possess biochemical properties favorable to in vitro applications, such as improved processivity and fidelity. Through homology and functional screens we have identified numerous viral replicases and some associated accessory proteins (e.g. helicases and primases), which are currently being developed as reagents that show promise to improve speed, throughput, accuracy, and reliability of DNA and RNA analysis. These DNA polymerases (and reverse transcriptases) from thermal aquifers are highly divergent from known enzymes and some are completely unique classes of enzymes, with little or no homology to Taq or Pfu DNAP. Several have been expressed to produce enzymes with novel properties and unique utility. The most extensively studied replicase allows high-fidelity, high efficiency PCR amplification of otherwise refractory sequences. An inherent reverse transcriptase activity allows single-tube, single-enzyme RT-PCR detection of RNA. This enzyme also allows Sanger sequencing of templates refractory to analysis using existing enzymes. A new class of DNA polymerases and affinity modifications promises to improve a variety of next generation sequencing and amplification applications.

FF0071

Optimizing DNA Shearing for Next-generation Sequencing

Jim Laugharn, Paul Ventura, Hamid Khoja, Jennifer Wu

Covaris, Inc., 14 Gill Street, Unit H, Woburn, MA 01801

As the emerging next-gen DNA sequencing continues to broaden genomic applications to diverse areas in biology and biomedicine, the importance of a reproducible, non-biased random DNA shearing becomes more critical to the sequencing process.

We will demonstrate how to optimize the DNA shearing process using the Covaris AFA (Adaptive Focused Acoustics) technology. The optimization consists of choosing the right instrument settings, such as duty cycle, intensity, cycle per burst, processing time, and the right processing vessels. With the right settings, the Covaris process is highly reproducible, delivers tightly distributed DNA fragments throughout a broad range, from 100bp to over 5kb. The isothermal, non-contact processing in closed vessels results in high recovery and no cross contamination, which are critical for any sequencing applications downstream.

The Covaris method can easily be automated for higher throughput needs, as well as be integrated with other reagent processes to further improve the sequencing workflow. Covaris L8 instrument incorporates a “Line transducer” instead of “point transducer” for other instruments. This design enables L8 to process samples in parallel. In addition, the L8 uses a next generation acoustic circuitry with a greater dynamic range.

Preliminary data using Covaris L8 for DNA shearing application shows equivalent performance to our “point transducer” based instruments. In addition, the L8 is eight times faster than our previous instrument because eight (8) samples can be treated simultaneously in 96-well format. As the energy is a continuous line, the system may be used with 96, 384, 1536, and other density plates.

The L8 system is integration ready; presents plate to front of the apparatus in a similar manner as plate readers. With the integration, Covaris DNA shearing processes become part of the automated sample preparation workflow before sequencing analysis. The Covaris Process enables an industrial-style approach to sample prep for sequencing. In addition, the Covaris process enables true thermal control to eliminate the thermal-biased fragmentation that is intrinsically inherent with other technologies (e.g. probe sonicators).

NOTES

NOTES

Wine & Cheese Poster Session

430pm – 630pm, May 28th

Sponsored by OpGen, Inc.

Enjoy!!!

Poster Presentations (Odd #s, May 28th)

FF0003

454 Sequencing Services at SeqWright: From Design, to Sequencing and Analysis

Xin-Xing Tan, Brad Thomas, Clayton M. Morrison, and Fei Lu

SeqWright, 2575 W. Bellfort St., Suite 2001, Houston, TX 77054-5025

The 454 GS FLX system from Roche was the first next-generation (Next-Gen) DNA sequencing platform to achieve commercial production and features a unique mix of long reads, exceptional accuracy, and ultra-high throughput. This system produces reads which are ten times the length of competitive next generation platforms, making this system ideal for whole genome de novo sequencing as well as targeted resequencing projects. As with other Next-Gen platforms, the FLX system can generate vast amounts of data per run within a short period of time. However, given the lack of familiarity most researchers have with Next-Gen platforms, the complex variety of potential Next-Gen applications and the overwhelming amounts of data theses platforms generate, consultative assistance and bioinformatic support is critical to the successful provision of Next-Gen sequencing services.

At SeqWright, we work with researchers closely before, during, and after a sequencing project, in order to help them make the right choice between the long-reads and lower throughput of the GS FLX and the short-reads and higher throughput of Applied Biosystems SOLiD, on which SeqWright also provides service. In this manner, SeqWright is capable of developing cost-effective and scientifically sound strategies for getting the most information out of the data generated. We have also developed an extensive IT infrastructure along with a number of bioinformatics tools which enable us to store, process, and analyze the massive sequencing datasets generated by Next-Gen platforms. All together, we are capable of providing a robust and flexible Next-Gen service portfolio which makes genomics projects of any size and scope accessible to the average researcher. Our Next-Gen services have been supporting a wide variety of project applications including whole-genome sequencing/resequencing, SNP discovery, metagenomics, transcriptome sequencing, non-coding RNA sequencing and ChIP sequencing.

FF0013

A Software System for Validating and Orienting Sequence Contigs Using Optical Maps

Adam Briska, Mihai Pop, Niranjan Nagarajan

OpGen, Inc., Gaithersburg, MD.

Although next generation sequencing methods are capable of rapidly producing high volumes of low-cost sequence data, the process of sequence finishing remains a significant challenge, due in part to the difficult task of determining the relative positions and orientations of sequence contigs, which is further complicated by the fact that sequence contigs can often contain mis-assemblies. Physical maps are powerful aids to resolving these complexities, but traditional BAC mapping can be expensive and time-consuming. Optical Mapping, on the other hand, is a technology which rapidly produces low-cost high-resolution ordered restriction maps of the entire genome independent of the sequence.

MapSolver is a graphical software tool which orients and validates sequence contigs against an Optical Map within minutes. MapSolver loads sequence information in from FASTA format and converts each sequence contig into an in silico Optical Map by finding the locations of restriction sites. Then, using a dynamic programming algorithm, it finds the optimal alignments of the in silico maps with the Optical Map. Based upon these alignments, MapSolver produces a visual representation of the relative orientations of these contigs and clearly shows any mis-assembled contigs. Additionally, this same information is easily exported into a tabular report.

This combination of the Optical Map data and MapSolver software can drastically reduce finishing times and costs.

FF0021

Comparative Sequence Analysis for Predictive Assembly Gap Closure

Andrey Kislyuk

Georgia Institute of Technology

Scaffolding and gap closing are critical steps in shotgun sequencing, and the increasing accessibility of new sequencing technologies results in a strong pressure to make these processes more affordable. A model of large-scale mutation events is proposed to allow contig layout and prediction of sequence in the gaps. Based on this model, the proposed gap closure algorithm forms a list of all putative gaps from multiple whole-genome alignment data, then analyzes conserved element structure in the neighborhood of each gap to produce an estimated probability that the given gap is the site of an insertion, deletion, inversion, or rearrangement mutation. It estimates the gap size and predicts the sequence present in the gap along with a confidence value for the prediction.

FF0031

Fungal -omics of an extreme environment promise transformational lignocellulolytic technologies and provide insight into the effects of key global environmental change drivers on ecosystem function

Amy Powell

Sandia National Laboratory, Albuquerque,NM

The global shift from fossil fuels to biofuels will require transformational breakthroughs in biomass deconstruction technologies, because current biofuel methods are neither cost effective nor sufficiently efficient or robust for scaleable production. Characterization of lignocellulolytic enzyme systems adapted to extreme environments will accelerate progress. Obvious extreme environments to mine for novel lignocellulolytic deconstruction technologies include aridland ecosystems (ALEs), such as those of the Sevilleta Long Term Ecological Research (LTER) site in central New Mexico. ALEs represent at least 35% of the terrestrial biosphere and are classic extreme environments, wherein low nutrient availability, high UV flux, limited and erratic precipitation, and extreme variation in temperatures represent the prevailing abiotic conditions. ALEs are functionally distinct from mesic environments in many respects; one salient distinction is that ALEs do not accumulate soil organic carbon (SOC), in marked contrast to mesic settings, which typically have large pools of SOC. Low productivity ALEs do not accumulate carbon (C) primarily because of extraordinarily efficient extracellular enzyme activities (EEAs) that are derived from underlying communities of diverse, largely uncharacterized microbes. Such efficient enzyme activities presumably reflect adaptation to this low productivity ecosystem, with the result that all available organic nutrients are assimilated rapidly. These communities are dominated by ascomycetous fungi, both in terms of abundance and contribution to ecosystem-scale metabolic processes, such as nitrogen (N) and C cycling. To deliver novel, robust, efficient lignocellulolytic enzyme systems that will drive transformational advances in biomass deconstruction, we propose a eukaryotic microbial metatranscriptomic inventory of blue grama grass (Bouteloua gracilis) rhizosphere (RHZ) soils. Blue grama is a keystone species in this and other North American grassland ecosystems. By sampling experimental plots with manipulated temperature and N, the proposed metagenomic inventory will also address the impacts of factors being evaluated in ongoing ecosystem studies that examine the effects of increased anthropogenic N deposition and climate change. Results obtained in the proposed metatranscriptomic inventory shall deliver novel, robust, efficient lignocellulolytic enzyme systems that will transform the biomass deconstruction technological landscape. At the same time, our work will give functional insight into community-level responses to the separate and combined effects of global-environmental change drivers.
Scope of Work:

Samples will be prepared from eukaryotic microbes that are endemic to blue grama RHZ soils. We request 600 Mb of high throughput sequence, representing replicated samples across four experimental treatments. We will require support from the JGI for expressed sequences clustering, functional annotation and human validation of predictions. We also seek collaboration with JGI in the areas of analyses and interpretation. Annotation jamborees that include key members of the fungal and environmental –omics communities, as well as, participants from the recently established DoE BioEnergy centers, are requested. Additional sequencing beyond the proposed 600 Mb may be required, should significant problems with clustering and annotation of expressed sequences become evident.

FF0039

Genome Improvement Pipeline Supporting The Human Microbiome Program

Margaret Priest, Harindra M. Arachchi, Amr Abouelleil, Marc Kelechava, Rakela Lubonja, Daniel Bessette, Michael G FitzGerald

Broad Institute/MIT

Next generation sequencing platforms have fueled a dramatic increase in our ability to generate draft genome assemblies. It is now routine to generate high quality bacterial draft assemblies based on data from the 454 Life Sciences platform. Generation of finished genome sequence has not experienced the same acceleration. An international working group has generated a framework that partly seeks to rebalance this growing divide. A series of five different sequence submission categories are now available, each describing a different state between draft and completely finished. The Human Microbiome Program seeks to generate approximately 1000 reference genome sequences for organisms associated with the human body. We will describe the Broad Institute's approach to maximize the information provided to the scientific community by performing limited improvement on all of our HMP genomes, while graduating a more novel group of genomes to higher levels of improvement, up to fully finished. We will describe our assembly improvement and finishing pipeline and provide preliminary data on HMP genomes moving through this pipeline.

FF0053

Genome Improvement with Bubble-PCR and Roche/454 Reads

Hope N. Tice, Janey Lee, Alicia Clum, Alla Lapidus, Alex Copeland, and Jan-Fang Cheng

US Department of Energy Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, California 94598 USA

As Sanger sequencing is being replaced by higher throughput and lower cost of the second-generation sequencing, finishing of microbial genomes and eukaryotic genome improvement will face two major challenges. First, the technology will need to be fast enough to handle many more drafted genomes.

Second, it will have to incorporate a clone-free approach to fill gaps. The method that utilizes a universal “bubble-tag” to perform primer walking and gap closure in a clone-free condition has been implemented at the Joint Genome Institute (JGI) and JGI-LANL using the Sanger based sequencing. The “bubble-tag” method was first described by Doug Smith (PCR Methods Appl. 2: 21-27, 1992) to amplify and sequence lambda DNA. We have started experimenting the bubble-PCR approach using sequences from the Roche/454 platform. Here we describe the experimentation of this approach in primer walking of the finished microbes, Micrococcus luteus NCTC 2665, and Methanococcus voltae A3. This test is to see if there is a sequencing bias toward high GC or low GC microbes. We also describe this experimentation method on a fungal project Aspergillus carbonarius. Genomic DNA was sheared, blunt-end repaired, or digested with frequent cutters, and ligated to bubble adaptors. Site specific primers were used together with the universal bubble primer to amplify and sequence the regions of interest. Roche/454 libraries were generated from the pooled amplification products. This approach enables primer walking and gap filling in a clone-free draft sequencing process. As described in the past the uniformity of this approach is amenable for an automated finishing/genome improvement process.

FF0055

Using 454 and Illumina Sequencing for Pseudogene Analysis in Three *Salmonella enterica* Strains by Comparison with High Quality Finished Genome Sequences

Craig Corton and the Pathogen Sequencing Unit

Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SA, UK

At the Wellcome Trust Sanger Institute we are developing strategies for the rapid comparison of bacterial genomes sequenced on 454 and Illumina GAII platforms against reference genomes. As part of our Salmonella research programme we are studying the evolutionary history of 3 serovars by genome comparisons against the finished sequences of non-host specific Salmonella enterica serovar Enteritidis PT4 and chicken specific S. enterica serovar Gallinarum 287/91. These reference sequences were determined by dye terminator chemistry on ABI3700 automated sequencers and were assembled, finished and fully annotated (Thomson et al. 2008). The genome sequences of S. enterica serovar Dublin BA207 and SC50 (isolated from cattle) and S. enterica serovar Pullorum 449/87 (a poultry specific serovar) were obtained by high coverage shotguns on the 454/Roche FLX platform. The contigs from Newbler assemblies were ordered and orientated against the finished genome of S. Enteritidis PT4 using an in-house automatic contig ordering script (ABACAS). The gene annotations from the finished strain were transferred to the newly sequenced genomes and further automated gene predictions were applied to the genomes to capture additional coding sequences (CDS) novel to each strain. The gene predictions were shown in the Artemis Comparison Tool (ACT) and this was used to identify which of the known pseudogenes of S. Gallinarum were also pseudogenes in the S. Dublin and S. Pullorum genomes. All other CDSs transferred over from the reference genome containing one or more mutations rendering the genes inactive were also identified. These regions will be confirmed by Illumina sequencing or by sequencing PCR products. A strategy for the rapid and accurate identification of pseudogenes provides a method for determining the evolutionary histories in Salmonella enterica serovars and provides information on their adaptation to specific hosts. The development and application of this strategy will be presented.

FF0071

Optimizing DNA Shearing for Next-generation Sequencing

Jim Laugharn, Paul Ventura, Hamid Khoja, Jennifer Wu

Covaris, Inc., 14 Gill Street, Unit H, Woburn, MA 01801

As the emerging next-gen DNA sequencing continues to broaden genomic applications to diverse areas in biology and biomedicine, the importance of a reproducible, non-biased random DNA shearing becomes more critical to the sequencing process.

We will demonstrate how to optimize the DNA shearing process using the Covaris AFA (Adaptive Focused Acoustics) technology. The optimization consists of choosing the right instrument settings, such as duty cycle, intensity, cycle per burst, processing time, and the right processing vessels. With the right settings, the Covaris process is highly reproducible, delivers tightly distributed DNA fragments throughout a broad range, from 100bp to over 5kb. The isothermal, non-contact processing in closed vessels results in high recovery and no cross contamination, which are critical for any sequencing applications downstream.

The Covaris method can easily be automated for higher throughput needs, as well as be integrated with other reagent processes to further improve the sequencing workflow. Covaris L8 instrument incorporates a “Line transducer” instead of “point transducer” for other instruments. This design enables L8 to process samples in parallel. In addition, the L8 uses a next generation acoustic circuitry with a greater dynamic range.

Preliminary data using Covaris L8 for DNA shearing application shows equivalent performance to our “point transducer” based instruments. In addition, the L8 is eight times faster than our previous instrument because eight (8) samples can be treated simultaneously in 96-well format. As the energy is a continuous line, the system may be used with 96, 384, 1536, and other density plates.

The L8 system is integration ready; presents plate to front of the apparatus in a similar manner as plate readers. With the integration, Covaris DNA shearing processes become part of the automated sample preparation workflow before sequencing analysis. The Covaris Process enables an industrial-style approach to sample prep for sequencing. In addition, the Covaris process enables true thermal control to eliminate the thermal-biased fragmentation that is intrinsically inherent with other technologies (e.g. probe sonicators).

FF0085

Bacterial Genome Assembly Validation using Optical Restriction Map

Dibyendu Kumar and William Farmerie

Interdisciplinary Center for Biotechnology Research,
University of Florida, Gainesville, FL-32611

High throughput next-generation DNA sequencers such as 454 Life Sciences GS FLX Titanium and ABI SOLiD have certainly revolutionized our approach to DNA sequencing. However, these sequencing techniques due to its in-vitro nature and short read length also bring new problems in genome assembly and finishing. Genome finisher has a tough time in arranging contigs or validating assembly in absence of physical map. Here we are emphasizing the importance of optical map to validate bacterial genome sequencing. An optical map is a single complete genome restriction map derived from a number of partial restriction fragment maps information. Basically, whole-genome optical maps are ordered restriction maps generated by spreading whole chromosomes onto treated glass surfaces containing many channels, followed by its digestion with restriction enzymes. About 50–100 contiguous restriction fragment of size measuring up to one-third of the whole chromosome are selected. These overlapping partial chromosome contigs are combined by alignment software using contiguous fragment sizes. The contiguous fragments of one optical map now aligned and compared to the in silico chromosome map of a sequenced reference strain. The optical map allowed us to identify several assembly errors, which is not possible without any mapping data. Despite the advantages of so-called ‘long sequence’ (~250 bp) pyrosequencing reads and clone end-pairing data, 454 assembly contains error because of the presence of numerous highly repetitive sequences. We, thus conclude that, in order to ensure the accuracy of finished genome sequences optical mapping is an important tool to validate de-novo assemblies generated by next-generation sequencers.

FF0095

Decontamination of MDA reagents for single cell genomics

Damon Tighe, Tanja Woyke, Janey Lee, and Jan-Fang Cheng

US Department of Energy Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, California 94598 USA

Single cell genomics, the sequencing of genomes from single cells, provides a glimpse into the genetic make-up and thus life style of the vast majority of uncultured microbial cells, making it an immensely powerful and increasingly popular tool. While DNA-free reagents for the amplification of a single cell genome are a prerequisite for successful single cell sequencing and analysis, DNA contamination has been detected in various reagents, which poses a considerable challenge. In this study we aimed to test the effect of UV radiation in reducing or eliminating the availability of the contaminated DNA for MDA. Exposure of the MDA buffer and enzyme to UV radiation prior to the amplification can reduce the amount of contaminating DNA in the end products, and this effect increases with an increasing amount of UV radiation. Although the enzymatic activity of phi-29 is also affected by the UV treatment, the remaining activity seems to have no problem in amplifying the target DNA to a similar level of DNA quantity. The methodology is quick, simple, and highly efficient and exhibits a low impact on the MDA reaction efficiency

FF0103

Adapting AUTOFINISH for Next-Generation Sequencing Technologies

Matt Reardon

NCGR, Santa Fe, NM

Finishing is a manual process requiring manpower in researchers and lab personnel as well as many different types of software and database tools which traditionally have not been integrated into a single efficient pipeline. In addition, next generation sequencing technologies have introduced new challenges with shorter reads and multiplexing issues.

NextGen AutoFinish is a pipeline developed to streamline the automated finishing of microbial and small eukaryotic organisms. It involves targeting finishing features such as intra-scaffold gaps, low coverage regions and repetitive areas for automatic primer design and preparation for laboratory reactions. As such, it remains an integral part of the overall high-throughput pipeline at JCVI. To keep up with new developments in sequencing technology, NextGen AutoFinish has been re-engineered to make the pipeline more robust and receptive to change going forward.

By utilizing a common data format, we have removed dependency on current chemistry or assembly technique. We have enhanced the ability of the end user to track reactions with integration into our award winning JLIMS system. We track targeted features in a hierarchical format so that they are easily query-able and provide reporting capability as to success and re-targeting during subsequent rounds. New feature targets are easily integrated as technology and project needs change. We have also leveraged automated assembly modules that exploit co-location constraints placed on finishing reads with Celera Assembler.

Future developments will include leverage multiple strain information, reference based projects, and ,as well as possible, integration with optical map data. We are also committed to making the project open source and supporting new features so other researchers/labs/centers can use the tools.

We present our experience applying this pipeline to de novo whole-genome assembly done with next generation sequencing.

FF0107

The Human Reference Genome, V 2.0

Tina Graves for the Genome Reference Consortium

The Wellcome Trust Sanger Institute, The Genome Center at Washington University, EBI and NCBI (http://genomereference.org)

The description of the essentially complete human genome sequence described by the International Human Genome Sequencing Consortium in 2004 marked a scientific milestone[2]. At the time, it was the only mammalian assembly of such high quality and coverage. While the assembly described in this publication has been the basis of much scientific discovery, such discovery has also made it clear that additional work is needed to allow the genome representation to grow as our understanding of the human genome grows. To this end, the Genome Reference Consortium (GRC) was created. Over the past two years, we have built systems and tools to facilitate the assembly and curation of clone based assemblies such as the reference human and mouse genomes. The culmination of this effort has resulted in an updated human genome assembly, referred to as Genome Reference Consortium Human Build 37 (GRC H37). We will review the significant improvements in this new human reference assembly. The initial focus of the group has been to fix assembly and sequence errors identified in human Build 36. Notably, the availability of new clone resources in non-BAC vectors[3] has also allowed us to re-examine gaps previously thought to be inaccessible to cloning. We have been able to add sequence to or completely close approximately one third of the unspanned gaps in Build 36. Additionally, much of the sequence that had been considered part of the 'bottom-drawer' has now either been localized or determined to be redundant and removed from the assembly. To accommodate our growing awareness of genetic diversity, this assembly also provides alternate representations of highly divergent loci for a limited number of regions. Additionally, the GRC has increased access to the data underlying the genome assembly. Tools to review component overlaps, as well as curation decisions, are publicly available. Regions that are currently under review can be seen at both the GRC portal and Ensembl Browser. Users can also report problems and ask questions through the GRC website.

[2]Nature (2004) **431**:931-45, [3]Nature (2008) **453**:56-64

FF0133

Screening of Recessive Genetic Disorders by Next Generation Sequencing

Darrell L. Dinwiddie, Callum Bell, Stephen F. Kingsmore

National Center for Genome Resources, Santa Fe, NM

Human recessive genetic diseases are individually quite rare, but taken together they are a major medical burden that cause significant morbidity and mortality. Among children, 20-30% of all infant deaths and 11% of pediatric hospital admissions are for children with genetic disorders. For example, Juvenile Batten Disease is a progressive, fatal neurodegenerative disorder of childhood caused by accumulation of lipopigment in neuronal tissue. Although rare, Juvenile Batten disease affects several hundred children in the United States and is only one of thousands of similar rare genetic disorders. In collaboration with the Beyond Batten Disease Foundation, a non-profit organization established to eradicate Batten disease through education and treatment, NCGR is developing a screening test that will be able to test for several hundred rare genetic disorders caused by thousands of mutations. Genomic DNA will be enriched for the genes of interest using target selection and then subjected to multiplexed, deep-sequencing on the Illumina GAIIx. Carrier status will be identified bioinformatically by aligning sequencing reads to a custom reference sequence and the subsequent identification of the presence or absence of known mutations. This screening test will be low cost, flexible, and available by 2010.

Funding: Beyond Batten Disease Foundation

FF0145

Simulating Assembler Performance: a Useful Preliminary Step to de novo Assembly

Guillaume Barreau, Marcin Swiatek, Thodoros Topaloglou

McGill University and Genome Quebec Innovation Center

Various software tools are available for the de novo assembly of massively parallel sequencers (MPS). However, given the difficulty of the task, none is capable of solving the problem completely. For any particular project, there is no information available on which of those tools is likely to yield the best results.

Faced with this choice for assembling the 454 reads of a lactobacillus strain, we tried to build an objective measure of the confidence we could have in the various assemblers available rather than choose one based on reputation alone. Our approach can be described as follows. We searched the literature for an indication of relatedness between various lactobacilli species and looked for one that would be close to ours and for which a complete genome had been submitted to a public database. This led us to Lactobacillus Plantarum whose genome is available on NCBI (NC_004567.1 GI:28376974).

From this sequence, we generated short dna fragments as a simulation of the shearing process. We generated reads of constant size (400 bp to simulate titanium data), perfectly tiled over the genome with a uniform coverage of 20x (one read starting every 20 base) and without any error. This meant that the data was unrealistically clean and made the problem significantly simpler for the assembler. However, we felt that this would still reveal which assembler was best for this kind of genome.

The fragments were then fed to the following assemblers: newbler, celera (wgs), mira2 and clc-bio. The resulting assemblies where then evaluated for correctness against the correct known answer to the problem collecting the following data for each assembler: the ratio of correct large contigs (>500 bp) produced and the proportion of bases from the original genome that were included in one of those contigs. On this particular genome, the best results were obtained with newbler and celera which both assembled 98.8% of the genome
correctly in 27 and 23 contigs respectively.

The relatively high number of contigs obtained gave us a background against which to measure the success of our real data assembly as well as guiding our choice of tools. We feel that this is a simple yet valuable step to introduce in an assembly project. We intend to release some simple software to automate the process and the collection of such data.

FF0165

Community-involved Microbial Genome Analysis at JGI-LANL

Susana Delano, Jean Challacombe, Gary Xie, Monica Misra, Thomas Brettin, Chris Detter

Los Alamos National Laboratory, Los Alamos, NM

The role of the genome analysis team at JGI-LANL is to facilitate publication of JGI genome papers and provide bioinformatics support and training to promote community-involved genome analysis. Our team members work closely with JGI collaborators on the comparative analysis of their genomes, with the goal of publishing the results in high profile scientific journals. When a JGI collaborator needs help with analysis and paper preparation, we assign an analysis team member to their project, in either a leading or supporting role. If we are leading the effort, we design and do most of the analysis, write the paper, and take the first author position on the paper. If we support the collaborator's effort, we perform specialized analyses, but someone from the collaborator's lab takes the lead on the paper. We also offer bioinformatics training for those collaborators who want to take the lead on the analysis and paper writing effort. Since 2005, we have conducted on-site, week-long, bioinformatics training sessions for 10 JGI collaborators, students and postdocs. Over the past 4 years, we have hosted visits by 36 JGI collaborators as part of our Genomic Explorers Seminar Series. We routinely receive requests for specialized analyses from JGI collaborator labs. In projects where JGI-LANL team members played a leading role in the analysis and preparation of genome papers, more than 10 genome papers have been published and 4 manuscripts are in preparation. We are currently expanding our efforts in the areas of single cell and metagenomic analysis.

This work was performed under the auspices of the US Department of Energy's Office of Science, Biological and Environmental Research Program, and by the University of California, Lawrence Berkeley National Laboratory under contract No. DE-AC02-05CH11231, Lawrence Livermore National Laboratory under Contract No. DE-AC52-07NA27344, and Los Alamos National Laboratory under contract No.DE-AC02-06NA25396.

Poster Session Notes

Poster Session Notes

Poster Session Notes

05/29/2009 - Friday				
Time	**Type**	**Abstract #**	**Title**	**Speaker**
730 - 830am	Breakfast	x	**Healthy Start Breakfast Buffet** (Scrambled Eggs on side tomatoes, scallions and spinach, Turkey sausage links, Assorted chilled fruit juices, Platter of freshly sliced seasonal fruit, Assorted and bran muffins with butter, Granola and oatmeal served with low-fat milk, Individual assorted fruit yogurts, etc.)	x
830 - 845	Intro	x	Welcome Back	**Chris Detter**
x	Session Chair	x	Session Chair	Chair – Alla Lapidus
845 - 930	**Keynote**	**FF0088**	**A Common Framework for Multiple Sources of Bacterial Annotation**	**Owen White**
930 – 950	Speaker 1	FF0005	BioHDF: Toward Scalable Bioinformatics Infrastructures	**Todd Smith**
950 – 1010	Speaker 2	FF0008	Using Consed and Cross_match in Resequencing Projects	**David Gordon**
1010 – 1030	Break	x	**Beverages and snacks provided**	x
1030 – 1050	Speaker 3	FF0126	Medicago truncatula Resequencing of 384 Lines	**Joann Mudge**
1050 – 1110	Speaker 4	FF0084	Performance Comparison of Multiple Genome Partitioning Technologies	**Jon Armstrong**
1110 – 1130	Speaker 5	FF0141	The IMG Systems for Comparative Analysis and Annotation of Microbial Genomes and Metagenomes	**Victor Markowitz**
1130 – 1150	Speaker 6	FF0022	GenePRIMP: Improving Microbial Gene Prediction Quality	**Amrita Pati**
1150 – 1210	Speaker 7	FF0137	Automated Microbial Genome Annotation: the Current State and Future Challenges	**Miriam Land**
1210 - 1230	Closing Discussions	x	**Closing Discussions -** discuss next year's plans	Chair - Chris Detter
1230 - 200pm	Lunch & Close of meeting	x	**La Fiesta Plaza Lunch Buffet -** (Chicken and beef fajitas with grilled red onions and bell peppers, Black beans (Vegetarian), Spanish rice (Vegetarian), Pork posole & calabacitas rancheras, Warm flour tortillas & butter, etc.) **End of meeting, enjoy lunch and Santa Fe**	**Sponsored by illumina**

Speaker Presentations (May 29th)

Abstracts are in order of presentation according to Agenda

FF0088

Keynote

Owen White

A Common Framework for Multiple Sources of Bacterial Annotation

University of Maryland

FF0005

BioHDF: Toward Scalable Bioinformatics Infrastructures

Todd M Smith, Eric Olson, Mark Welsh

Geospiza, Inc. Seattle WA USA (www.geospiza.com)

"If the data problem is not addressed, ABI's SOLiD, 454's GS FLX, Illumina's GAII or any of the other deep sequencing platforms will be destined to sit in their air-conditioned rooms like a
Stradivarius without a bow" was the closing statement in the lead Nature Biotechnology editorial" Prepare for the deluge" (Oct. 2008). The oft-stated challenges focus on the obvious problems of storing and analyzing data. However, the problems are much deeper than the short descriptions portray. True, researchers are ill-prepared to confront the challenges of inadequate IT infrastructures, but there is a greater challenge in that there is a lack of easy to use, well performing software systems and interfaces that would allow to researchers to work with data in multiple ways to summarize information and drill down into supporting details.

Meeting the above challenge requires that we have well performing software frameworks and
underlying data management tools to store and organize data in better ways than complex mixtures of flat files and relational databases. Geospiza and The HDF Group are collaborating to develop open-source, portable, scalable, bioinformatics technologies based on HDF5 (Hierarchical Data Format – http://www.hdfgroup.org). We call these extensible domain-specific data technologies "BioHDF." BioHDF will implement a data model that supports primary DNA sequence information (reads, quality values, meta data) and the results from sequence alignment and variation detection algorithms. BioHDF will extend HDF5 data structures and library routines with new features (indexes, additional compression, graph layouts) to support the high performance data storage and computation requirements of Next Gen Sequencing.

For close to 20 years, HDF data formats and software infrastructure have been used to manage and access high volume complex data in hundreds of applications, from flight testing to global climate research. The BioHDF effort is leveraging these strengths. We will show data from small RNA and gene expression analyses that demonstrate HDF5's value for reducing the space, time, bandwidth, and development costs associated with working with Next Gen Sequence data.

FF0008

Using Consed and Cross_match in Resequencing Projects

David Gordon and Phil Green

Genome Sciences Dept, Univ of Washington and Howard Hughes Medical Institute

We have adapted the sequence editor Consed and the sequence comparison program Cross_match for use in resequencing projects with large numbers of short reads. Cross_match can now find gapped alignments for a million 36bp Solexa reads against the human genome in about 1.5 hours (about 18 minutes if cross_match is allowed to ignore matches within masked repeats). Consed-ready ace files can be prepared in a fraction more time.

Cross_match reports, for resequencing applications, sequence variants & confirmed segments and strength of evidence. It has many options for reporting of hits (including mapping qualities) and can report a histogram of hit scores. It can find local, global (the entire read is aligned), or RNASeq alignments (allowing cDNA to be aligned against genomic DNA with GY-AG splice sites).

Consed now has the capability to read and display Solexa and 454 reads (and their "traces"), and can handle resequencing assemblies with ~5 million aligned reads with response times similar to those for small assemblies. Additional features have been added to Consed, such as navigators for variants and high/low depth of coverage regions, and sorting of the reads within a window, making it useful (and not overwhelming) to view such assemblies.

We have implemented CALF, a compact file format for both unaligned reads and alignments with the goal of facilitating nimble analysis and visualization of alignments with up to billions of reads. CALF files of unaligned reads are roughly half the size of fastq files with the same data.

These and other improvements will be presented.

FF0126

Medicago truncatula Resequencing of 384 Lines

Joann Mudge1, Michael J. Sadowsky2, Peter L. Tiffin2, Maria J. Harrison3, Betsy M. Martinez-Vaz4 and Gregory D. May1, Nevin D. Young2

1 National Center for Genome Resources, Santa Fe, NM 87505
2 University of Minnesota, St. Paul, MN 55108
3 Boyce Thompson Institute, Ithaca, NY 14853
4 Hamline University, St. Paul, MN 55014

Medicago truncatula is a genomic model for the plant legume family, important for its ability to fix nitrogen through symbiotic relationships with microorganisms. We are currently resequencing nearly 400 diverse M. truncatula lines using solexa sequencing technology. This will allow a genome-wide survey of SNPs, indels, and structural variants through alignments of each resequenced genome to the M. truncatula genome reference assembly. A subset of thirty lines will be sequenced deeply (30X) allowing detection of most genomic variants and the creation of a core set of variation within M. truncatula. The rest of the lines will be skim sequenced (5X). Genomic variation will be catalogued and compared among lines. Shared ancestral genomic segments (haplotypes) will be identified providing a basis for genome-wide association mapping. The sequenced lines will also be phenotyped in symbiotic relationships. Correlations of symbiosis phenotypes with genotypes or haplotypes will uncover genomic segments important in symbiotic relationships. All variants and haplotypes found in this study will be made available to the community to enable further studies of genomic regions underlying important legume traits.

FF0084

Performance Comparison of Multiple Genome Partitioning Technologies

Jon R. Armstrong, Vincent Magrini, Ryan Demeter, Daniel C. Koboldt, Richard K. Wilson, and Elaine R. Mardis

The Genome Center at Washington University School of Medicine, St. Louis, MO

The ability and capacity to investigate human genetic variation is a major goal now that the human genome is sequenced. Massively parallel sequencing technologies take us one step closer to assessing the genetic variation between individuals and diseases. However, considering the complexity and repeat content of the human genome, there is an important need for technologies that target and extract a defined subset of the genome for variation discovery.

Historically, PCR has been the predominant procedure for the selection and enrichment of a relevant genomic region. However, PCR is labor intensive, expensive at large scale, and failure-prone. Further, the complexity of the human genome, coupled with the limitations on PCR product size reduces its utility in large-scale studies. New technologies are being developed to help mitigate some of these problems by specifically targeting and isolating, in a parallel fashion, multiple regions of the human genome for resequencing. These include oligonucleotides on microarrays or biotinylated probes produced from oligonucleotides that are designed to hybridize the regions of interest. Both technologies have the advantage that they target thousands of areas in the human genome but may suffer from variable sequence coverage uniformity and questionable hybridization specificity with a concomitant decrease in the ability to detect variants.

To this end, data was generated by hybridizing multiple CEPH DNA samples from the 1000 Genomes Project to several targeted capture platforms followed by sequencing on the Illumina or Roche instruments. We will present results that analyze and assess the performance of several genome partitioning technologies with respect to depth and breadth of coverage of targeted areas, hybridization specificity of the targeted probes, and variant detection capability.

NOTES

FF0141

The IMG Systems for Comparative Analysis and Annotation of Microbial Genomes and Metagenomes

Victor M. Markowitz1 and Nikos C. Kyrpides2

1Biological Data Management and Technology Center, Lawrence Berkeley National Laboratory, USA
2Genome Biology Program, Department of Energy Joint Genome Institute, USA

Microbial genome and microbial community metagenome analysis are growing areas that are expected to lead to advances in healthcare, environmental cleanup, agriculture, industrial processes, and alternative energy production. With the rapid growth in the number of microbial genome and microbial community metagenome sequence datasets, comparative data analysis plays a critical role in understanding the biology of newly sequenced organisms and communities. The effectiveness of comparative analysis depends on the availability of an integrated genome data context, powerful analytical tools, the quality of genome annotations, and the level of detail in cellular reconstruction. Analyzing genome and metagenome datasets jointly with other (e.g., phylogenetically related) datasets is substantially more efficient than analyzing each dataset in isolation.

FF0022

GenePRIMP: Improving Microbial Gene Prediction Quality

Amrita Pati, Natalia Ivanova, Sean D. Hooper and Nikos C. Kyrpides

DOE Joint Genome Institute, Genome Biology Program,
2800 Mitchell Drive, Walnut Creek, CA

Post-Sanger high-throughput sequencing technologies such as 454, SOLiD, and SOLEXA have revolutionized DNA sequencing by facilitating sequencing on a scale that was previously thought to be prohibitive. This positive growth by many orders of magnitude of the amount of DNA that can be sequenced also poses challenges in the form of new kinds of sequencing anomalies affecting downstream analysis and the necessity of computational pipelines that can detect and correct such anomalies, while maintaining efficiency with the high volumes of data.

We present GenePRIMP (Gene PRediction IMprovement Pipeline), a computational pipeline that evaluates the accuracy of gene models in genomes/metagenomes at different stages of finishing. GenePRIMP identifies anomalies in gene models such as inconsistent start codons, missed genes, and frameshift-based errors, some of which are results of errors in finishing. Such anomalies, besides being an indicator of the finishing quality of the genome, can also reveal and compare the accuracies of prevalent genome annotation methods. GenePRIMP is available as a web-based application at http://geneprimp.jgi-psf.org/, as well as a stand-alone application on request.

FF0137

Automated Microbial Genome Annotation: the current state and future challenges

Miriam Land, Loren Hauser, Frank Larimer, Yun-Juan (Janet) Chang, Cynthia Jefferies, Gwo-Liang Chen, and Bob Cottingham

Oakridge National Laboratory, Oakridge, TN

JGI has sequenced numerous bacterial genomes with the goal of furthering DOE's missions such as bioremediation, carbon sequestration, and energy production. New advances in sequencing technologies will greatly accelerate the rate of sequencing. Manual genome annotation is would be ideal but it does not scale with the increased rate of DNA sequence generation. ORNL has built an automated microbial genome annotation pipeline (http://genome.ornl.gov/microbial/) that provides JGI with accurate protein coding gene models, functional descriptions of those protein products, preliminary biochemical pathway categorization, structural RNA gene models, some regulatory RNA models, and CRISPR repeat identifications. The pipeline's multiple tools and database queries are used to create a reference web site for each genome, and are used as the basis for both automated and manual annotation. New tools are under development to provide controlled and consistent nomenclature for various functional categories of proteins. The first of these tools has been incorporated into the pipeline, and identifies and categorizes most of the regulatory proteins in an organism. A second tool for the identification and categorization of transporters is almost ready to be incorporated into the pipeline. Finally, additional tools that build upon our existing suite of tools are needed to extract additional useful biological information and increase the quality and speed of the automated annotation.

NOTES

Discussion Notes

Discussion Notes

Lunch

12:30 – 2:00pm

Sponsored by

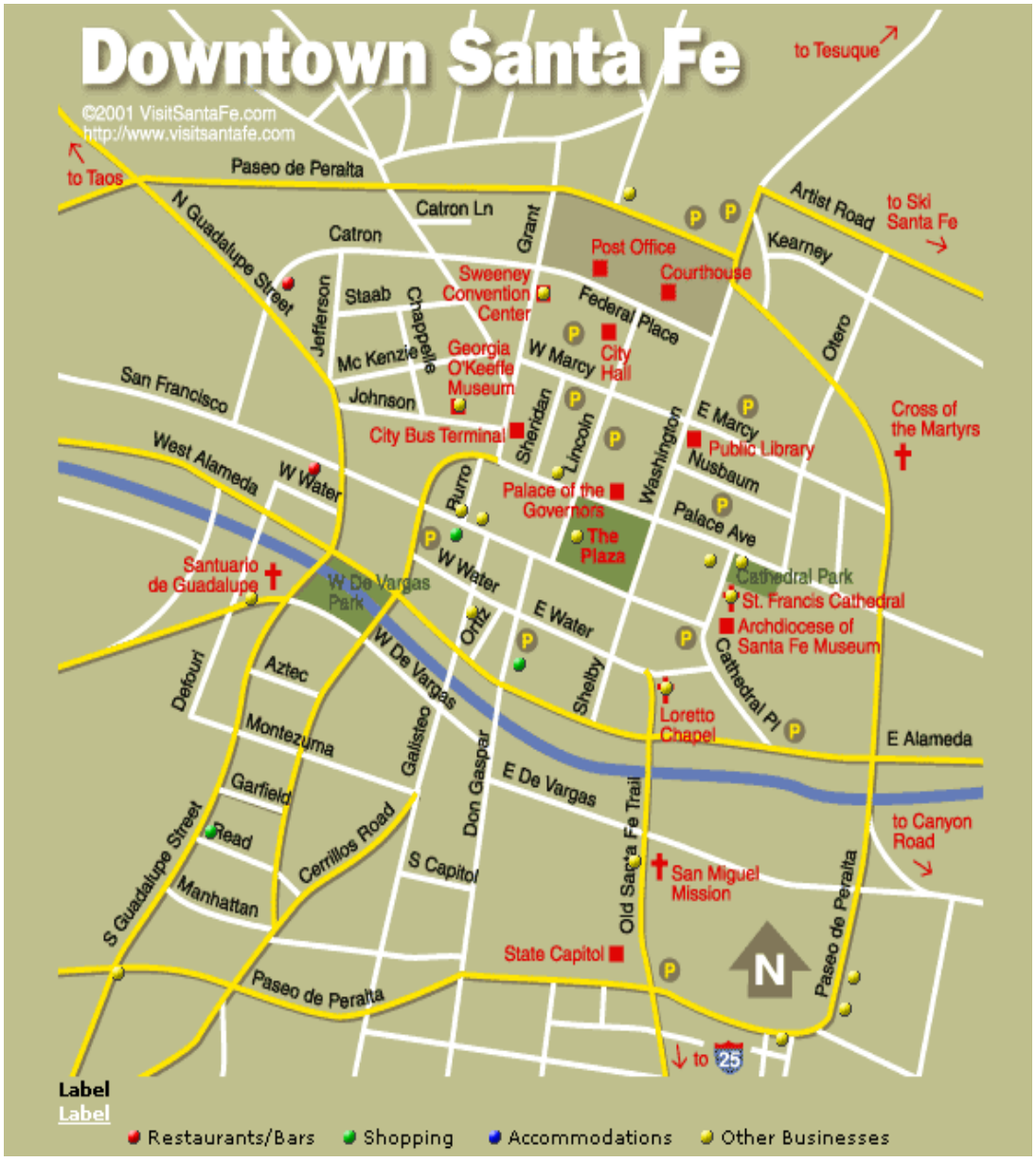
2009 Attendee List

FF #	Name	Affiliation	email
1	**Chris Detter**	Los Alamos National Laboratory - JGI	cdetter@lanl.gov
2	**Jane Hutchinson**	Roche Applied Science	jane.hutchinson@roche.com
3	**Xin-Xing Tan**	SeqWright	xxtan@seqwright.com
4	**Mansi Chovatia**	Joint Genome Institute - LBL	mrchovatia@lbl.gov
5	**Todd Smith**	Geospiza, Inc	todd@geospiza.com
6	**Robert Blakesley**	NIH Intramural Sequencing Center (NISC)	rblakesl@nhgri.nih.gov
7	**Ernie Retzel**	National Center for Genome Resources (NCGR)	efr@ncgr.org
8	**David Gordon**	Howard Hughes Medical Institute at University of Washington	dgordon@u.washington.edu
9	**Omayma Al-Awar**	Edge BioSystems	oalawar@edgebio.com
10	**Stephen Kingsmore**	National Center for Genome Resources (NCGR)	sfk@ncgr.org
11	**Greg May**	National Center for Genome Resources (NCGR)	gdm@ncgr.org
12	**Teri Mueller**	Roche Diagnostics	teri.mueller@roche.com
13	**Adam Briska**	OpGen, Inc.	abriska@opgen.com
14	**John Crow**	National Center for Genome Resources	jac@ncgr.org
15	**Tim Hunkapiller**	Discovery Bio	tim@discoverybio.com
16	**Valérie Barbe**	GENOSCOPE	vbarbe@genoscope.cns.fr
17	**Sophie Mangenot**	GENOSCOPE	mangenot@genoscope.cns.fr
18	**Christophe Battail**	GENOSCOPE	cbattail@genoscope.cns.fr
19	**Evan Skowronski**	ECBC	evan.skowronski@us.army.mil
20	**Julia Scheerer**	Tauri - TMTI	Julia.Scheerer_CONTRACTOR@dtra.mil
21	**Andrey Kislyuk**	Georgia Institute of Technology	kislyuk@gatech.edu
22	**Amrita Pati**	LBL	apati@lbl.gov
23	**Isaac Meek**	Caliper Life Sciences	Isaac.Meek@caliperls.com
24	**Niranjan Nagarajan**	University of Marylanc	niranjan@umiacs.umd.edu
25	**Steve Turner**	Pacific Biosciences	sturner@pacificbiosciences.com, trard@pacificbiosciences.com
26	**Jessica Hostetler**	J Craig Venter Institute (JCVI)	Jessicah@jcvi.org
27	**Darren Grafham**	The Wellcome Trust Sanger Institute	dg1@sanger.ac.uk
28	**John Havens**	Integrated DNA Technologies	jhavens@idtdna.com
29	**Keven Stevens**	Integrated DNA Technologies	x
30	**Ken Taylor**	Integrated DNA Technologies	ktaylor@idtdna.com
31	**Amy Powell**	Sandia National Laboratory	ajpowel@sandia.gov
32	**Wes Warren**	Washington University in St. Louis	wwarren@watson.wustl.edu
33	**Sofi Ibrahim**	USAMRIID	sofi.ibrahim@us.army.mil
34	**Andrew Stewart**	Naval Medicial Research Center	Andrew.Stewart@med.navy.mil
35	**Shanmuga Sozhamannan**	Naval Medicial Research Center	Shanmuga.Sozhamannan@med.navy.mil
36	**Gary Roemer**	New Mexico State University	groemer@nmsu.edu
37	**Brook Milligan**	New Mexico State University	brook@biology.nmsu.edu
38	**Tim Harkins**	Roche Diagnostics	tim.harkins@roche.com
39	**Margaret Priest**	Broad Institute/MIT	mpriest@broad.mit.edu
40	**Olga Chertkov**	Los Alamos National Laboratory - JGI	ochrtkv@lanl.gov
41	**Karen Davenport**	Los Alamos National Laboratory - JGI	kwdavenport@lanl.gov
42	**Michael Rey**	Novozymes. Inc.	MWR@novozymes.com
43	**Martha Trela**	Pacific Biosciences	mtrela@pacificbiosciences.com
44	**Alfredo Lopez De Leon**	Novozymes, Inc	ALLO@novozymes.com
45	**Randy Berka**	Novozymes, Inc	RAMB@novozymes.com
46	**Theresa Hepburn**	Broad Institute/MIT	thepburn@broad.mit.edu
47	**Sante Gnerre**	Broad Institute/MIT	sante@broad.mit.edu
48	**William Spencer**	OpGen, Inc.	wspencer@opgen.com
49	**Thomas Brettin**	Los Alamos National Laboratory - JGI	brettin@lanl.gov
50	**Gary Resnick**	Los Alamos National Laboratory	resnick@lanl.gov
51	**Christine Sun**	UC Berkeley	christine.l.sun@gmail.com
52	**Nicole Rosenzweig**	OptiMetrics, Inc.	nrosenzweig@omi.com
53	**Hope Tice**	Joint Genome Institute - LLNL	tice1@llnl.gov
54	**Danielle Walker**	The Wellcome Trust Sanger Institute	dw2@sanger.ac.uk
55	**Craig Corton**	The Wellcome Trust Sanger Institute	chc@sanger.ac.uk
56	**Sarah Young**	Broad Institute/MIT	stowey@broad.mit.edu
57	**Brian Thomas**	UC Berkeley	bcthomas@berkeley.edu
58	**Jim Bristow**	Joint Genome Institute - LBL	JBristow@lbl.gov
59	**Susan Lucas**	Joint Genome Institute - LLNL	lucas11@llnl.gov
60	**Tijana Glavinadelrio**	Joint Genome Institute - LLNL	glavinadelrio1@llnl.gov
61	**Marc Kelechava**	Broad Institute of MIT	marc@broad.mit.edu
62	**Anna Montmayeur**	The Broad Institute	annamont@broad.mit.edu
63	**Linda Meincke**	Los Alamos National Laboratory - JGI	meincke@lanl.gov
64	**Dave Klaasse**	DTRA - TMTI	david.klaasse@DTRA.MIL
65	**Mike Smith**	DTRA - CB	Michael.Smith@DTRA.MIL
66	**Alla Lapidus**	Joint Genome Institute - LBL	alapidus@lbl.gov
67	**Tao Zhang**	Joint Genome Institute - LBL	tzhang3@lbl.gov
68	**Kerrie Barry**	Joint Genome Institute - LBL	KWBarry@lbl.gov
69	**Hui Sun**	Joint Genome Institute - LBL	HSun@lbl.gov
70	**Kurt LaButti**	Joint Genome Institute - LBL	klabutti@lbl.gov
71	**James Laugharn**	Covaris, Inc.	JLaugharn@covarisinc.com

72	**David Bruce**	Los Alamos National Laboratory - JGI	dbruce@lanl.gov
73	**Lynne Goodwin**	Los Alamos National Laboratory - JGI	lynneg@lanl.gov
74	**Clayton Morrison**	Seqwright	cmorrison@seqwright.com
75	**Chris Munk**	Los Alamos National Laboratory - JGI	cmunk@lanl.gov
76	**Alicia Clum**	Joint Genome Institute - LBL	aclum@lbl.gov
77	**Liz Saunders**	Los Alamos National Laboratory - JGI	ehs@lanl.gov
78	**Johar Ali**	Ontario Institute for Cancer Research (OICR)	johar.ali@oicr.on.ca
79	**Jose Olivares**	Los Alamos National Laboratory	olivares@lanl.gov
80	**Steve Lowry**	Joint Genome Institute - LBL	slowry@lbl.gov
81	**Cliff Han**	Los Alamos National Laboratory - JGI	han_cliff@lanl.gov
82	**Sarah Pelan**	The Wellcome Trust Sanger Institute	sb2@sanger.ac.uk
83	**Hajnalka Kiss**	Los Alamos National Laboratory - JGI	hajkis@lanl.gov
84	**Jon Armstrong**	Washington University in St. Louis	jarmstro@watson.wustl.edu
85	**Dibyendu Kumar**	University Of Florida	dkumar@ufl.edu
86	**Eric Green**	NIH - NHGRI	egreen@nhgri.nih.gov
87	**Evan Eichler**	University of Washington	eee@gs.washington.edu
88	**Owen White**	University of Maryland	owhite@som.umaryland.edu
89	**Joel Martin**	Joint Genome Institute - LBL	j_martin@lbl.gov
90	**Jan Kieleczawa**	Wyeth Research	JKieleczawa@wyeth.com
91	**Mike Fitzgerald**	Broad Institute of MIT	fitz@broad.mit.edu
92	**Christian Buhay**	Baylor College of Medicine	cbuhay@bcm.tmc.edu
93	**Michael Rhodes**	Applied Biosystems	rhodesmd@appliedbiosystems.com
94	**Hilary Morrison**	MBL	morrison@mbl.edu
95	**Damon Tighe**	Joint Genome Institute - LBL	DJTighe@lbl.gov
96	**David Mead**	Lucigen Corp.	dmead@lucigen.com
97	**Shannon P Dugan**	Baylor College of Medicine	sdugan@bcm.tmc.edu
98	**Yan Ding**	Baylor College of Medicine	yding@bcm.tmc.edu
99	**Donna Muzny**	Baylor College of Medicine	donnam@bcm.tmc.edu
100	**Jason Miller**	J. Craig Venter Institute	jmiller@jcvi.org
101	**Scott Sammons**	Center for Disease Control (CDC)	zno6@CDC.GOV
102	**Michael Holder**	Baylor College of Medicine	mholder@bcm.tmc.edu
103	**Matt Reardon**	J. Craig Venter Institute	MReardon@jcvi.org
104	**Vincent Magrini**	Washington University in St. Louis	vmagrini@watson.wustl.edu
105	**Bob Fulton**	Washington University in St. Louis	bfulton@watson.wustl.edu
106	**Chad Tomlinson**	Washington University in St. Louis	ctomlins@watson.wustl.edu
107	**Tina Graves**	Washington University in St. Louis	tgraves@watson.wustl.edu
108	**Aye Wollam**	Washington University in St. Louis	awollam@wustl.edu
109	**Marcella Putman**	Applied Biosystems	Marcella.Putman@appliedbiosystems.com
110	**Jarret Glasscock**	Cofactor Genomics	jarret_glasscock@cofactorgenomics.com
111	**Patrick Minx**	Washington University in St. Louis	pminx@watson.wustl.edu
112	**Luke Tallon**	University of Maryland	ljtallon@som.umaryland.edu
113	**Kristie Jones**	University of Maryland	x
114	**Lori Peterson**	Caldera Pharmaceuticals, Inc	court@cpsci.com
115	**Nicole Touchet**	Caldera Pharmaceuticals, Inc	touchet@cpsci.com
116	**Ryan Kim**	National Center for Genome Resources (NCGR)	rwk@ncgr.org
117	**Marvin Stodolsky**	Dept. of Energy, OBER	Marvin.Stodolsky@science.doe.gov
118	**Matt Hickenbotham**	Cofactor Genomics	matt_hickenbotham@cofactorgenomics.com
119	**Neil Miller**	National Center for Genome Resources (NCGR)	nam@ncgr.org
120	**Leonda Clendenen**	National Center for Genome Resources (NCGR)	lec@ncgr.org
121	**Jimmy Woodward**	National Center for Genome Resources (NCGR)	jew@ncgr.org
122	**Faye Schilkey**	National Center for Genome Resources (NCGR)	fds@ncgr.org
123	**Joe Salvatore**	CLC bio	jsalvatore@clcbio.com
124	**Patrick Chain**	DOE Joint Genome Institute - LLNL	chain2@llnl.gov
125	**Deanna Church**	National Center for Biotechnology Information (NCBI)	church@ncbi.nlm.nih.gov
126	**Joann Mudge**	National Center for Genome Resources (NCGR)	jm@ncgr.org
127	**Craig Pierson**	Illumina, Inc.	cpierson@illumina.com
128	**Eric Mathur**	Synthetic Genomics, Inc.	EMathur@SyntheticGenomics.com
129	**Stephan Trong**	Joint Genome Institute - LLNL	trong1@llnl.gov
130	**Sergey Koren**	J. Craig Venter Institute	skoren@jcvi.org
131	**Jeremy Schmutz**	Hudson Alpha	jschmutz@hudsonalpha.org
132	**Haley Fiske**	Illumina, Inc.	hfiske@illumina.com
133	**Darrell Dinwiddie**	National Center for Genome Resources (NCGR)	dld@ncgr.org
134	**Robert Huffman**	Defense Threat Reduction Agency	Robert.Huffman@dtra.mil
135	**Haofeng Chen**	New Mexico State University	brook@nmsu.edu
136	**Alexander Tchourbanov**	New Mexico State University	brook@nmsu.edu
137	**Miriam Land**	Oak Ridge National Lab - JGI	landml@ornl.gov
138	**Peter Houde**	New Mexico State University	phoude@nmsu.edu
139	**Dan Drell**	Dept. of Energy, OBER	Daniel.Drell@science.doe.gov
140	**Jim Knight**	Roche Diagnostics - 454	james.knight@roche.com
141	**Victor Markowitz**	DOE Joint Genome Institute - LBL	vmmarkowitz@lbl.gov
142	**Jon Murray**	CLC bio	jmurray@clcbio.com
143	**David Sims**	Hudson Alpha	dsims@lanl.gov
144	**GongXin Yu**	Boise State University	gongxinyu@boisestate.edu
145	**Guillaume Barreau**	McGill University and Genome Quebec Innovation Center	guillaume.barreau@mail.mcgill.ca

146	**Patrice Milos**	Helicos BioSciences Corporation	PMilos@helicosbio.com
147	**Steve Lombardi**	Helicos BioSciences Corporation	slombardi@helicosbio.com
148	**Wendy Castle**	New Mexico State University	w.l.castle@gmail.com
149	**Alba Chavez**	New Mexico State University	x
150	**Zhong Wang**	Joint Genome Institute - LBL	ZhongWang@lbl.gov
151	**Feng Chen**	Joint Genome Institute - LBL	FChen@lbl.gov
152	**Cristina Vesbach**	University of New Mexico	cvesbach@unm.edu
153	**Bryce Ricken**	Sandia National Laboratory	bricken@sandia.gov
154	**Mike FitzPatrick**	OpGen, Inc.	mfitzpatrick@opgen.com
155	**Bob Cottingham**	Oak Ridge National Lab - JGI	cottinghamrw@ornl.gov
156	**Larry Kedes**	Institue for Genetic Medicine	laurence.kedes@keck.usc.edu
157	**David Watson**	Airforce Research Laboratory	David.Watson@WPAFB.AF.MIL
158	**Keith Brown**	RainDance Technologies, Inc.	brownk@raindancetech.com
159	**Take Ogawa**	RainDance Technologies, Inc.	lambertj@raindancetech.com
160	**Kenneth Frey**	Airforce Research Laboratory	Kenneth.Frey@WPAFB.AF.MIL
161	**John J. Schlager**	Airforce Research Laboratory	John.Schlager@WPAFB.AF.MIL
162	**Rafal Woycicki**	Warsaw University of Life Sciences - SGGW	rafalwoycicki@gmail.com
163	**Raymond Cologna**	Noblis	Raymond.Cologna@noblis.org
164	**Mithu Chatterjee**	University Of Florida	cmithu@ufl.edu
165	**Susana Delano**	Los Alamos National Laboratory - JGI	sdelano@lanl.gov

Map of Santa Fe, NM



History of Santa Fe, NM

Thirteen years before Plymouth Colony was settled by the Mayflower Pilgrims, Santa Fe, New Mexico, was established with a small cluster of European type dwellings. It would soon become the seat of power for the Spanish Empire north of the Rio Grande. Santa Fe is the oldest capital city in North America and the oldest European community west of the Mississippi.

While Santa Fe was inhabited on a very small scale in 1607, it was truly settled by the conquistador Don Pedro de Peralta in 1609-1610. Santa Fe is the site of both the oldest public building in America, the Palace of the Governors and the nation's oldest community celebration, the Santa Fe Fiesta, established in 1712 to commemorate the Spanish reconquest of New Mexico in the summer of 1692. Peralta and his men laid out the plan for Santa Fe at the base of the Sangre de Cristo Mountains on the site of the ancient Pueblo Indian ruin of Kaupoge, or "place of shell beads near the water."

The city has been the capital for the Spanish "Kingdom of New Mexico," the Mexican province of Nuevo Mejico, the American territory of New Mexico (which contained what is today Arizona and New Mexico) and since 1912 the state of New Mexico. Santa Fe, in fact, was the first foreign capital over taken by the United States, when in 1846 General Stephen Watts Kearny captured it during the Mexican-American War.

Santa Fe's history may be divided into six periods:

Preconquest and Founding
(circa 1050 to 1607)

Santa Fe's site was originally occupied by a number of Pueblo Indian villages with founding dates from between 1050 to 1150. Most archaeologists agree that these sites were abandoned 200 years before the Spanish arrived. There is little evidence of their remains in Santa Fe today.

The "Kingdom of New Mexico" was first claimed for the Spanish Crown by the conquistador Don Francisco Vasques de Coronado in 1540, 67 years before the founding of Santa Fe. Coronado and his men also discovered the Grand Canyon and the Great Plains on their New Mexico expedition.

Don Juan de Onate became the first Governor-General of New Mexico and established his capital in 1598 at San Juan Pueblo, 25 miles north of Santa Fe. When Onate retired, Don Pedro de Peralta was appointed Governor-General in 1609. One year later, he had moved the capital to present day Santa Fe.

Settlement Revolt & Reconquest
(1607 to 1692)

For a period of 70 years beginning the early 17th century, Spanish soldiers and officials, as well as Franciscan missionaries, sought to subjugate and convert the Pueblo Indians of the region. The indigenous population at the time was close to 100,000 people, who spoke nine basic languages and lived in an estimated 70 multi-storied adobe towns (pueblos), many of which exist today. In 1680, Pueblo Indians revolted against the estimated 2,500 Spanish colonists in New Mexico, killing 400 of them and driving the rest back into Mexico. The conquering Pueblos sacked Santa Fe and burned most of the buildings, except the Palace of the Governors. Pueblo Indians occupied Santa Fe until 1692, when Don Diego de Vargas reconquered the region and entered the capital city after a bloodless siege.

Established Spanish Empire
(1692 to 1821)

Santa Fe grew and prospered as a city. Spanish authorities and missionaries - under pressure from constant raids by nomadic Indians and often bloody wars with the Comanches, Apaches and Navajos-formed an alliance with Pueblo Indians and maintained a successful religious and civil policy of peaceful coexistence. The Spanish policy of closed empire also heavily influenced the lives of most Santa Feans during these years as trade was restricted to Americans, British and French.

The Mexican Period
(1821 to 1846)

When Mexico gained its independence from Spain, Santa Fe became the capital of the province of New Mexico. The Spanish policy of closed empire ended, and American trappers and traders moved into the region. William Becknell opened the l,000-mile-long Santa Fe Trail, leaving from Arrow Rock, Missouri, with 21 men and a pack train of goods. In those days, aggressive Yankeetraders used Santa Fe's Plaza as a stock corral. Americans found Santa Fe and New Mexico not as exotic as they'd thought. One traveler called the region the "Siberia of the Mexican Republic."

For a brief period in 1837, northern New Mexico farmers rebelled against Mexican rule, killed the provincial governor in what has been called the Chimayó Rebellion (named after a village north of Santa Fe) and occupied the capital. The insurrectionists were soon defeated, however, and three years later, Santa Fe was peaceful enough to see the first planting of cottonwood trees around the Plaza.

Territorial Period
(1846 to 1912)

On August 18, 1846, in the early period of the Mexican American War, an American army general, Stephen Watts Kearny, took Santa Fe and raised the American flag over the Plaza. Two years later, Mexico signed the Treaty of Guadalupe Hidalgo, ceding New Mexico and California to the United States.

In 1851, Jean B. Lamy, arrived in Santa Fe. Eighteen years later, he began construction of the Saint Francis Cathedral. Archbishop Lamy is the model for the leading character in Willa Cather's book, "Death Comes for the Archbishop."

For a few days in March 1863, the Confederate flag of General Henry Sibley flew over Santa Fe, until he was defeated by Union troops. With the arrival of the telegraph in 1868 and the coming of the Atchison, Topeka and the Santa Fe Railroad in 1880, Santa Fe and New Mexico underwent an economic revolution. Corruption in government, however, accompanied the growth, and President Rutherford B. Hayes appointed Lew Wallace as a territorial governor to "clean up New Mexico." Wallace did such a good job that Billy the Kid threatened to come up to Santa Fe and kill him. Thankfully, Billy failed and Wallace went on to finish his novel, "Ben Hur," while territorial Governor.

Statehood
(1912 to present)

When New Mexico gained statehood in 1912, many people were drawn to Santa Fe's dry climate as a cure for tuberculosis. The Museum of New Mexico had opened in 1909, and by 1917, its Museum of Fine Arts was built. The state museum's emphasis on local history and native culture did much to reinforce Santa Fe's image as an "exotic" city.

Throughout Santa Fe's long and varied history of conquest and frontier violence, the town has also been the region's seat of culture and civilization. Inhabitants have left a legacy of architecture and city planning that today makes Santa Fe the most significant historic city in the American West.

In 1926, the Old Santa Fe Association was established, in the words of its bylaws, "to preserve and maintain the ancient landmarks, historical structures and traditions of Old Santa Fe, to guide its growth and development in such a way as to sacrifice as little as possible of that unique charm born of age, tradition and environment, which are the priceless assets and heritage of Old Santa Fe."

Today, Santa Fe is recognized as one of the most intriguing urban environments in the nation, due largely to the city's preservation of historic buildings and a modern zoning code, passed in 1958, that mandates the city's distinctive Spanish-Pueblo style of architecture, based on the adobe (mud and straw) and wood construction of the past. Also preserved are the traditions of the city's rich cultural heritage which helps make Santa Fe one of the country's most diverse and fascinating places to visit.

IDT introduces the miRCat™ Cloning Kit for small RNA discovery

miRCat™ small RNA cloning is based on the pre-activated, adenylated linkering method that has been successfully used in many labs since its development in 2001[1]. miRCat™ permits cloning from any RNA source in any species.

Material sufficient for ten cloning experiments is provided in the miRCat™ Small RNA Cloning Kit, and a detailed technical manual provides instructions for cloning and sequencing small RNAs either as individual clones or as concatamers.

www.idtdna.com for more miRCat™ information

References
1. Lau NC, LP Lim, EG Wienstein, and DP Bartel 2001 An abundant class of tiny RNAs with probable regulatory roles in Caenorhabditis elegans. Science 294: 858-862.

IDT® INTEGRATED DNA TECHNOLOGIES

INNOVATION AND PRECISION IN NUCLEIC ACID SYNTHESIS

www.idtdna.com

ISO 9001:2000 FM88954 ISO 13485:2003 FM 513219

Go the Distance

ULTRAmers™

Ultra high-fidelity synthesis of oligos up to 200 bases with mass spec QC!

IDT continues to push the boundaries of DNA synthesis quality and performance. By leveraging our industry-leading synthesis platform and chemistry, we now offer the custom synthesis of Ultramers™, 60-200 base DNA oligos suitable for demanding applications such as cloning and gene construction. Save time and trouble with direct synthesis of the entire target fragment!

To maintain our commitment to high quality, a proprietary LC-MS method has also been developed to provide accurate mass assessment for Ultramers™. As always, this service is offered free of charge for each oligo with all data accessible online.

Ultra*mers*™ specifications:

- 60 to 200 DNA bases
- Delivered normalized and lyophilized in tubes
- Ideal for gene construction, cloning and ddRNAi
- PAGE purification and plate synthesis options also available

IDT®
INTEGRATED DNA TECHNOLOGIES

INNOVATION AND PRECISION IN NUCLEIC ACID SYNTHESIS

www.idtdna.com

ISO 9001:2000 FM88954
ISO 13485:2003 FM 513219

http://www.roche-diagnostics.us/
Meet and Greet Party

http://www.covarisinc.com/
Sport Duffle Bags

http://www.illumina.com/
Closing Lunch

http://www.opgen.com/
Wine and Cheese Poster Session

http://www3.appliedbiosystems.com/
Notebooks

http://www.idtdna.com/Home/Home.aspx
Meeting Guides